Social Media Sentiment, Investor Herding and Informational Efficiency

Ni Yang[a,*], Adrian Fernandez-Perez[b], Ivan Indriawan[c]

[a] *Auckland University of Technology*

[b] *University College Dublin*

[c] *University of Adelaide*

ABSTRACT

We examine the impact of social media sentiment on the informational efficiency of financial markets. Specifically, we explore the relationship between sentiment extracted from Twitter posts and two commonly used measures of efficiency: return autocorrelation and variance ratio. Our findings reveal that higher sentiment leads to higher return autocorrelation and variance ratio the following day, indicating a decrease in informational efficiency. We also demonstrate that the impact of social media sentiment on informational efficiency stems from the emergence of herding behaviors among traders, with higher sentiment leading to heightened herding activity. Our findings support the notion that higher social media sentiment contributes to a decline in the quality of the information environment, resulting in informationally inefficient equity prices.

JEL Classification: C31; G14; G41

Keywords: Social Media; Investor Sentiment; Market Quality; Informational Efficiency

Acknowledgment: We thank John Fan, Vitali Alekseev, Nuttawat Visaltanachoti, Fanqi Meng and participants at the 2023 Sydney Banking and Financial Stability Conference, the 2024 New Zealand Finance Colloquium, and the University of Otago Accounting and Finance seminar.

* Corresponding author: WF Building, 42 Wakefield Street, Auckland University of Technology, New Zealand, e-mail: ni.yang@aut.ac.nz

1. Introduction

Information plays a crucial role in allocating resources efficiently in the economy and financial markets. Informationally efficient prices arise when prices promptly and accurately integrate all publicly available information, and trading on such prices is expected to enhance asset allocation efficiency. The trading process hinges on market participants adjusting their beliefs and trading in response to new information arrivals. Numerous studies have examined the degree to which market is informationally efficient and have documented that markets have long-standing inefficiencies in reflecting new information. While Grossman (1995) argues that the asset allocation of trading process can generate noise signals, preventing markets from being informationally efficient, a strand of literature, starting with the seminal papers of Shiller (1981) and De Long et al. (1990), considers that investors are not rational as assumed and hence, are affected by sentiment and create noise in prices to news, which can mitigate information efficiency.

In the context of acquiring new information, investors rely on public news to update their knowledge and make informed investment decisions (De Long et al., 1990). Nowadays, however, social media has become the dominant source of information dissemination (Gan et al., 2020). While social media facilitates interactions among individuals and connects investors with financial markets, it can also lead to collective investment behaviors among market participants (Bukovina, 2016). As a result, investor sentiment becomes intertwined with the quality of a market, causing stock return continuations, increasing market frictions and potentially affecting the efficiency of asset prices. In this paper, we study how sentiment extracted from Twitter (now rebranded as X) posts impacts price informational efficiency.

From a behavioral standpoint, investors are not perfectly rational. Their investment decisions can be influenced by various factors, including their own mood, market sentiment, and other seemingly irrelevant external factors.[1] Interactions on social media platforms can alter the information environment of individuals and generate cycles of responses, potentially leading to sentimental hype. These impacts can subsequently affect investors' trading behaviors, asset prices, and the overall market efficiency. However, whether social media sentiment increases or decreases market efficiency remains unclear. To the best of our knowledge, this study is the first to directly investigate the relationship between social media sentiment and informational efficiency, while also delving into the underlying mechanisms involved.

Previous studies suggest a likely relationship between sentiment and informational efficiency. On the one hand, sentiment has the potential to enhance efficiency. For instance, Vozlyublennaia (2014) demonstrates that increased investor attention, measured through Google searches, reduces return predictability, and therefore, improves informational efficiency. Gu and Kurov (2020) show that social media sentiment extracted from Twitter posts provides new information about analyst recommendations, analyst price targets and quarterly earnings. Given that social media sentiment can convey firm-level information, it is reasonable to expect that it may contribute to enhancing informational efficiency.

On the other hand, social media sentiment may disrupt informational efficiency, especially in a high-frequency trading environment. For example, Da et al. (2011) highlight that investors may not effectively utilize their information sets due to variations in their ability to process new information. This effect could be more pronounced in high-frequency trading environments where investors have limited time to react and cope with rapid influx of social media messages. Additionally, social media sentiment disseminates information to a wider range of audiences, which can collectively foster irrational behaviors among investors in the

[1] Previous studies examine how social mood (Nofsinger, 2005), weather (Hirshleifer and Shumway, 2003), sporting events (Edmans et al., 2007), and music choices (Edmans et al., 2022) affect the stock markets.

stock market, such as herding (Li et al., 2023), overactions (Jiao et al., 2020) and irrationality to surprises (Karampatsas et al., 2023). These factors collectively contribute to the potential for irrational investment decisions that deviate from fundamental principles, ultimately diminishing market efficiency.

In this study, we employ a textual analysis approach to extract sentiment from social media content and investigate its impact on informational efficiency at high frequency. We focus on the aggregated tone of Twitter posts, commonly referred to as ‘tweets’, as a proxy for social media sentiment. To measure informational efficiency, we employ two commonly used metrics: return autocorrelation and variance ratio, consistent with previous studies such as Hendershott and Jones (2005), O’Hara and Ye (2011), and Comerton-Forde and Putniņš (2015). We regress these metrics on the sentiment measure to explore whether increased sentiment leads to changes in information efficiency. Our findings demonstrate that as social media sentiment increases, there is an increased return autocorrelation and variance ratio, indicating a decrease in informational efficiency. We account for various influential market factors, employ different sentiment analysis approaches, and consider different intervals for sentiment construction, all of which lend support to the robustness of our finding.

The current study also delves into the underlying mechanism for the above finding through the role of herding behavior. It is important to note that certain market participants have access to professionally curated reports and commercial databases that provide real-time trading data, enabling them to extract information from the trading activities of others. For other participants, however, these resources may be inaccessible or come at a high cost, leading them to rely more heavily on information obtained through other sources such as social media (Bukovina, 2016). This reliance on social media can result in herding behavior, potentially impacting informational efficiency. We consider two herding behavior metrics: dollar-based herding (Cai et al., 2019) and the Williams Percent Range (Zhou, 2018). Utilizing a vector autoregressive (VAR) model, we show that a higher social media sentiment leads to heightened herding activity, but not the inverse relationship. This finding complements Da et al. (2011) and Shen et al. (2017) who show that higher sentiment leads to higher trading frictions, and eventually slowing down the market information incorporation process.

Our study relates to literature on the impact of social media sentiment on informational efficiency, expanding studies such as Kurov (2008) and Vozlyublennaia (2014). We use intraday data to construct the informational efficiency measures and synchronous real time investor sentiment measure. Unlike the low frequency survey data used to proxy for investor sentiment in Kurov (2008), the granularity of intraday data enables in-depth and more accurate analysis of market dynamics, offering better insights on market efficiency and investors' reactions to news. [2] Our empirical analysis reveals that these impacts exhibit distinct characteristics, particularly in a high-frequency setting. Our study also explores the mechanism through which social media sentiment influences market informational efficiency. We demonstrate that, driven by social media, investors collectively engage in herding activities. This can have a detrimental effect on market efficiency, in line with previous studies such as Kumar and Lee (2006) and Barber et al. (2008). For instance, Kumar and Lee (2006) show that as individuals trade in the same direction as others, retail sentiment can trigger stock return co-movements. Barber et al. (2008) document that individual investors herd and their trades forecast future returns.

Our study has important implications for various stakeholders. Firstly, we provide evidence that social media sentiment has a substantial impact on the quality of equity markets, beyond the influence of other conventional market-based factors. This finding holds significant

[2] It is well-known that survey data has some shortages such as answering bias and lagged information due to long data collection process and low update frequency. Intraday data, on the other hand, allows for textual analytic sentiment to be matched with real-time market price dynamics. Thus, sentiment extraction using intraday data overcomes the non-synchronicity issue and answering bias, and is more suitable to study the impact of sentiment on informational efficiency.

relevance for market participants, indicating that they should consider social media sentiment as a crucial factor when formulating investment strategies. Secondly, for regulators and policymakers, our study highlights the potential of social media as an additional surveillance tool within the market regulatory framework. Recognizing the influence of social media sentiment can aid in enhancing market oversight to effectively monitor and manage potential risks and market disruptions.

The remainder of the paper proceeds as follows. In Section 2, we provide an overview of the relevant literature. Section 3 outlines the data used in our analysis and elaborates the construction of our variables of interest. Section 4 presents the results on the linkage between social media sentiment and informational efficiency. In Section 5, we explore the transmission channel underlying such linkage. Section 6 performs robustness tests. Section 7 concludes.

2. Literature review

Numerous studies have highlighted the significance of investors' behavior and reaction to news. While there is a group of 'smart' investors and high-frequency traders who can exploit the arrival of news (see, e.g., Busse and Green, 2002; Grinblatt et al., 2012; and Foucault et al., 2016), most market participants are not equipped with such processing skills. These investors collectively exhibit irrational reactions to news, resulting in less efficient prices. For instance, investors' underreaction to new information can result in short-term stock price continuation, indicating market inefficiency (Zhang, 2006). Moreover, De Bondt and Thaler (1985) and Tetlock (2011) show that investors' overreaction to news contributes to price deviations from fundamentals, and therefore, market inefficiency.

Despite these insights, there is a limited amount of research directly examining the relationship between social media sentiment, investors behavior and market informational efficiency. While social media has become a dominant channel of information sharing in recent years, existing studies predominantly focus on the role of social media sentiment in predicting returns. For instance, Chen et al. (2014) find that social media opinions are a significant source for future stock returns and earnings surprises predictions. The consensus among these studies is that a high sentiment is contemporaneously associated with positive returns, followed by a subsequent correction.[3] Bollen et al. (2011), for instance, demonstrate that incorporating social media sentiment significantly improves their model's predictive power on the Dow Jones Industrial Average (DJIA) index. More predictable returns indicate a potential negative impact of social media sentiment on market efficiency. Similarly, Kim et al. (2014) document that incorporating investor sentiment enhances profitability, which is indicative of reduced market efficiency. Furthermore, Duz Tan and Tas (2021) discover that social media sentiment predicts future returns even after controlling for news sentiment, implying that social media activity contains unique information beyond traditional news sources.

This linkage between social media sentiment and market efficiency can be particularly strong in shorter time horizons. This can be attributed to the slow reaction time of retail investors (De Long et al., 1990). Supporting this notion, Sun et al. (2016) discover that lagged half-hour SPY ETF investor sentiment can predict subsequent intraday S&P 500 index returns. Their findings demonstrate that social media sentiment holds economic value, exhibits distinctions from intraday momentum effects, and has a lasting impact. De Jong et al. (2017) demonstrate that the lagged innovation of tweets impacts the returns of 87% of the stocks in the DJIA at the minute level, alleviating concerns about sentiment's impact being limited to specific stocks or markets. Guégan and Renault (2021) further support these findings by documenting that pricing efficiency in cryptocurrency markets decreases as the frequency

[3] Investor sentiment has significant predictive power for US stock returns (Baker and Wurgler, 2006) and other markets such as Canada, France, Germany, Japan and the UK (Baker et al., 2012). Siganos (2014) documents that Facebook Gross National Happiness Index positively predicts following day stock market returns, but with a partial price reversal over the following weeks.

increases, indicating heightened market inefficiencies at shorter horizons. Collectively, this evidence reinforces the influence of social media sentiment on market informational efficiency.

Furthermore, research has highlighted the association between increased sentiment and feedback trading, a form of investor herding, which in turn is linked to greater return predictability and market inefficiency.[4] For instance, Kurov (2008) studies feedback trading using E-mini S&P 500 and E-mini Nasdaq-100 data. Using weekly survey data as a proxy for investor sentiment, he finds that positive feedback trading appears to be more active in periods of high investor sentiment. Similarly, Chau et al. (2011) observe a connection between Baker and Wurgler's (2006) investor sentiment index and the returns of three major ETFs (S&P 500 ETF Trust, Dow Jones Industrial Average ETF Trust, and the Invesco QQQ ETF). They demonstrate that optimistic (pessimistic) investors are more (less) likely to adopt trend-chasing investment strategies at the daily level. Conversely, however, Kaplanski and Levy (2014) document that sophisticated investors can exploit sentiment and restore efficiency in the US market. These studies indicate that social media sentiment may play a role on market efficiency through the collective imbalanced orders of irrational traders, contributing to herding behavior.

Recent studies have shed light on the role of sentiment-driven herding behaviors in explaining anomalies such as abnormal returns observed during periods of extremely high sentiment in US and European markets (Filip and Pochea, 2023). Through a causality test, Blasco et al. (2012) find that sentiment and past returns drive herding behaviors among investors, and buyer (seller)-initiated herding is more pronounced when past returns are positive (negative). As individuals tend to follow the same sign of orders than others, retail sentiment can trigger stock return co-movements (Kumar and Lee, 2006). This observation is

[4] Positive feedback trading is a strategy which buys when prices move up and sell when prices move down. Such a strategy may be due to behavioral biases on the part of some investors. In the presence of positive feedback trading, it may be optimal for rational speculators to jump on the bandwagon. The interaction between feedback traders and rational speculators moves prices away from fundamentals in the short run (De Long et al., 1990).

further corroborated by Barber et al. (2008) and Da et al. (2011), who document that increased investor attention can lead to higher trading volume and abnormal returns due to net buying pressure of retail investors. However, it is important to note that sentiment-driven irrational behaviors, such as feedback trading or herding, are not limited to retail investors alone. They are also observed among fund managers (Lakonishok et al., 1992; Menkhoff and Nikiforow, 2009), analysts (Welch, 2000; Clement and Tse, 2005), and institutional investors. For instance, Nofsinger and Sias (1999) show a positive correlation between herding and lagged returns among institutional investors, with the effect being even stronger than in individual investors.

Interactions among investors on social media platforms have emerged as a potential avenue for investor herding, as discussed in Fenzl and Pelzmann (2012). The extensive user engagement on platforms like Twitter, involving sharing and responding to news and messages related to stocks, leads to enhanced connectivity among investors and contributes to collective investment behaviors (Bukovina, 2016). As a result, market-wide herding behaviors can arise, influencing the net orders placed in the market and subsequently, harming market efficiency.

3. Data and measures of informational efficiency

3.1. Tweets and sentiment extraction

We focus on the SPDR S&P 500 ETF Trust (ticker: SPY) as a representation of the US equity market. We collect tweets using Twitter's official Application Programming Interface (API). Following Sprenger et al. (2014), we use cashtags ($) to search for tweets related to a particular security, i.e., '$SPY' to obtain tweets related to SPY. Our sample period is from August 1, 2012 to March 31, 2022 since Twitter only officially introduced cashtags on July 31, 2012. We collected 6.85 million tweets and every tweet is reported in the US Eastern Standard Time (EST) and time-stamped to the nearest second. Figure 1 plots the average number of tweets mentioning $SPY by the day of the week and by the hour of the day. The volume of tweets is significantly higher during trading days and, particularly, during trading hours between 9:30 and 16:00 EST. Thus, we focus on these periods for our analyses. We clean each tweet by removing irrelevant characters, including punctuations, emojis and internet links. These filters lead to a total of 2,433 trading days in consideration.

[Insert Figure 1 Here]

We use the WordNet lexical database as a language processing tool to transform qualitative into quantitative data. It was developed by the Cognitive Science Laboratory of Princeton University and has been widely adopted for social media sentiment evaluation and classification (see e.g., Navigli, 2009; Vidhu Bhala and Abirami, 2014; AlMousa et al., 2021). Using WordNet in natural language processing allows us to score each tweet between -1 and 1. We consider a tweet as positive if its score is greater than zero, negative if the score is less than zero, and neutral when the score is zero.[5]

We aggregate the directional tone from tweets to a daily level, which we then use to construct our social media sentiment index. Following studies such as Antweiler and Frank (2004), Sprenger et al. (2014), and Leung and Ton (2015), we construct our social media sentiment, $Sentiment_t$, as follows,

$$Sentiment_t = \ln\left[\frac{1+M_t^{Positive}}{1+M_t^{Negative}}\right], \tag{1}$$

where $M_t^{Positive}$ and $M_t^{Negative}$ are the sum of positive and negative tweets during market trading hours on day *t*, respectively. This measure captures the overall sentiment embedded in

[5] In addition to this sentiment classification method, we employ other methods, such as the Harvard IV-4 sentiment list (Tetlock, 2007), the Loughran-McDonald sentiment list (Loughran and McDonald, 2014), and SentiWordNet (Azar and Lo, 2016) in our robustness section.

tweets for each day. A high (low) $Sentiment$ reflects a more optimistic (pessimistic) view on the SPY.

Panel A of Table 1 reports the daily summary statistics for the social media sentiment. The sentiment on the SPY is positive, with an average value of 0.76. This is consistent with the existing literature which shows that investors are generally optimistic about the financial markets (Baker and Wurgler, 2006; Stambaugh et al., 2012).

[Insert Table 1 Here]

Figure 2 plots the five-day moving average sentiment (dotted line) and the daily SPY price (solid line) over the sample period from August 2012 to March 2022. The visualised two series display an intuitive co-movement between SPY prices and social media sentiment, motivating us to explore further regarding the role of sentiment on informational efficiency.

[Insert Figure 2 Here]

3.2. Stock market data

For our stock market data, we obtain transaction-level data of SPY from Refinitiv Tick History. The data contains all activity observed at the national best bid and ask, time-stamped to the nearest millisecond. We omit the first and last ten minutes of trading to avoid the confounding effects of market opening and closing. To minimize the effect of recording errors, we exclude transactions where trading volume is above the day's 99.9^{th} percentile. We then follow Chordia et al. (2001) and remove observations containing non-positive quoted spread, quoted spread greater than 5, effective spread/quoted spread greater than 4, percentage effective spread/percentage quoted spread greater than 4, and quoted spread/transaction price greater than 0.4.

For multiple trades that are executed with the same time-stamp, we treat them as one trade as they often reflect a trade initiated by one market participant but executed against the limit orders of multiple participants. In such cases, we use the value-weighted average transaction price and aggregate the volume traded. We then follow Lee and Ready (1991) trade signing algorithm to classify each trade into buyer- and seller-initiated trades. A trade is classified as buyer- (seller-) initiated if the transaction price is above (below) the prevailing midquote. For trades that occur at the midquote, we employ the tick rule and compare the current price with the previous. The construction of informational efficiency measures considered in this study requires price data at various frequencies. As such, we aggregate the transaction-level data to 1-, 10-, 30- and 60-sec intervals. Finally, we winsorize all the continuous series at the 1% each tail to reduce the effect of outliers.

3.3. Informational efficiency measures

We follow Comerton-Forde and Putniņš (2015) and construct two informational efficiency measures. These metrics measure the extent to which asset prices deviate from a random walk. First, we calculate the daily absolute midquote return autocorrelation at different frequencies. This metric gauges efficiency by capturing both the under and overreaction of returns to information arrival. Smaller values indicate that prices follow a random walk, and therefore, a more efficient market. The equation is defined below,

$$Autocorrelation_{t,k} = \left|Corr\left(r_{t,k,n}, r_{t,k,n-1}\right)\right|. \tag{2}$$

$r_{t,k,n}$ is the n^{th} midquote return measured at intraday frequency k for a given day t, where $k \in$ {1-sec, 10-sec, 30-sec}. Using the absolute values of autocorrelation across three different frequencies, we apply a principal component analysis (PCA) and extract the first principal component, $Autocorrelation^{PCA}$. We then re-scale it so that it ranges from zero (most efficient) to one (least efficient). As explained in Comerton-Forde and Putniņš (2015), the absolute autocorrelation at a single frequency contains some degree of measurement noise. The first principal component reduces this noise and, therefore, is a more accurate measure of efficiency.

The second informational efficiency measure is the absolute excess variance ratio. This measure indicates whether the relationship between the variance of returns at various horizons is linear. The underlying assumption for an efficient market is that the variance of its returns is equal to k times the variance measured at a higher frequency (Lo and MacKinlay, 1988). The equation is as follows,

$$VarianceRatio_{t,kl} = \left| \frac{\sigma^2_{t,kl}}{k\sigma^2_{t,l}} - 1 \right|, \quad (3)$$

where $\sigma^2_{t,l}$ and $\sigma^2_{t,kl}$ are the variance of l-second and k-second midquote return for a trading day t. We use different combinations for (l, kl), i.e., (10-sec, 30-sec), (10-sec, 60-sec) and (30-sec, 60-sec). Similar to the previous metric, we apply a PCA and extract the first principal component, $VarianceRatio^{PCA}$. A higher value indicates slower incorporation of information, and therefore, lower informational efficiency.

Panel A of Table 1 further reports the statistical summary of the market efficiency measures. The autocorrelation and variance ratios are, on average, 0.16 and 0.12, respectively, indicating some degree of informational inefficiency. For comparison, Frijns et al. (2023) report a cross-sectional mean of 0.093 for autocorrelation and 0.082 for variance ratio across the S&P 500 constituent stocks. Interestingly, the initial autocorrelation, denoted as AR(1), for the market efficiency metrics is notably modest, hovering around 0.06. This observation implies that instances of market inefficiency are promptly rectified, lacking any enduring impact over time. In the next section, we examine the relationship between social media sentiment and market informational efficiency.

4. Empirical results

4.1. Baseline specification

We assess the relationship between social media sentiment and informational efficiency. Our baseline model regresses the informational efficiency measures on the social media sentiment as follows,

$$Y_t = \alpha + \beta \cdot Sentiment_{t-1} + \delta \cdot Y_{t-1} + \gamma \cdot Controls_t + \varepsilon_t, \tag{4}$$

where Y_t is one of the two measures of informational efficiency on day t, i.e., $Autocorrelation^{PCA}$ or $VarianceRatio^{PCA}$. To avoid endogeneity issues, social media sentiment is lagged one day[6], $Sentiment_{t-1}$. The parameter of interest is β which reflects the impact of social media sentiment on informational efficiency. We include the lagged dependent variable, Y_{t-1}, to control for persistence in the informational efficiency metrics. $Controls_t$ are variables known to influence the patterns of return serial correlations, as highlighted by McKenzie and Faff (2003) and McKenzie and Kim (2007). These variables include the daily SPY return, realized volatility, dollar volume, average bid and ask depth, and the stock market implied volatility. The contemporaneous setup for these control variables is consistent with studies on market quality such as Hendershott et al. (2011) and Brogaard et al. (2015) while the motivation for using these controls is as follows. First, stock returns have been positively

[6] The causal relation between social media sentiment and market informational efficiency is the interest of this study, the lagged specification is consistent with the literature (Frijns et al., 2023).

associated with return autocorrelation.[7] Second, empirical evidence by Chau et al. (2011) indicates that realized volatility exerts a negative influence on serial correlations.[8] The feedback trading hypothesis suggests that increased volatility tends to reduce the presence of positive feedback traders, thereby mitigating autocorrelation. Third, heightened trading volume, often a reflection of more informed trading, bolsters market efficiency (McKenzie and Faff, 2003). Consequently, increased trading volume lowers the potential for short-term return predictability, fostering greater market efficiency. Fourth, the bid-ask depth is anticipated to amplify informational efficiency by integrating trading-induced price impacts and information into prevailing prices. Finally, we incorporate the S&P 500 implied volatility index (VIX) to account for overall market uncertainty, with the prevailing expectation of an inverse correlation between VIX and autocorrelation (or variance ratio). The daily SPY return and VIX are collected from Refinitiv Workspace and the CBOE, respectively. The remaining control variables are retrieved from Refinitiv Tick History.

We report the correlations between our variables in Panel B of Table 1. Sentiment has low but positive associations with both autocorrelation and variance ratio. This indicates that the market is less (more) efficient in optimistic (pessimistic) periods. The relationships between the informational efficiency metrics and the control variables are in line with our expectations. That is, return is positively correlated with autocorrelation and variance ratio, while other control variables negatively correlate with them.

Table 2 reports the regression estimates of Equation (4) with the autocorrelation as the dependent variable. Column (1) indicates that sentiment positively and significantly impacts the autocorrelation of SPY. A higher social media sentiment reduces informational efficiency

[7] Positive return autocorrelation is more frequently observed during a market upward trend, while negative return autocorrelation in more likely during market downturn (McKenzie and Faff, 2003). In addition, Valadkhani (2022) shows that prices of large ETF, such as SPY, increase more during market uptrend compared to the decrease during market downturn.

[8] We define the daily realized volatility as the square root of the sum of the squared SPY midquote returns at 1-minute frequency from 9:30 to 16:00 EST.

the following day. Columns (2) to (6) show that this effect persists after we include the control variables. In line with Table 1, the lagged autocorrelation has a positive and significant but small coefficient, suggesting that autocorrelation is not highly persistent. Moreover, the coefficients for the control variables are consistent with the expected sign discussed previously. For instance, return is positively associated with autocorrelation. Realized volatility and the VIX have a negative effect on autocorrelation, i.e., an improvement in informational efficiency. This can be explained using the feedback trading hypothesis where increased volatility reduces the number of positive feedback traders in the market. Higher dollar volume reduces autocorrelation, and accordingly, improves informational efficiency. Finally, the average bid-ask depth is negatively associated with autocorrelation.[9]

[Insert Table 2 Here]

Column (7) shows that the effect of social media sentiment on informational efficiency is robust to the inclusion of various controls. We observe that social media sentiment remains impactful on autocorrelation. A one standard deviation increase in sentiment is associated with a 0.009 higher autocorrelation (or a 7.2% increase in autocorrelation).[10] The findings indicate that investor behavior is not entirely rational and can be influenced by content shared on Twitter related to the SPY. Social media interactions may reshape investors' informational landscape, foster a collective enthusiasm within a market, and influence investment choices. Our results support the notion that investor sentiment interlaces with stock returns and causes pricing

[9] There are two explanations for this finding. First, the average bid-ask depth serves a direct indicator of order-induced price impact, a conduit for information-driven trading activities, as outlined in Hasbrouck (1991). Second, higher average bid-ask depth weakens the impact of bid-ask bounce (Roll, 1984), thereby contributing to a more subdued impact and an improved level of informational efficiency.

[10] This is calculated as $0.030 \times 0.31 = 0.009$ where the regression coefficient (0.03) is multiplied by the standard deviation of the sentiment index (0.31) reported in Table 1. This is equivalent to a $(0.030 \times 0.31) \div 0.13 = 7.2\%$ increase in autocorrelation, where 0.13 is the full sample standard deviation of autocorrelation shown in Table 1.

frictions, in line with the observation of Da et al. (2011). Our findings are also consistent with Shen et al. (2017), who ascertain that markets display greater irrationality and diminished efficiency during optimistic periods.

In Table 3, we report the regression estimates of Equation (4) with the variance ratio as the dependent variable. Consistent with the previous table on autocorrelation, we also find that sentiment has a negative impact on the variance ratio. A one standard deviation increase in sentiment is associated with a 0.005 higher variance ratio (or the 5.2% of its full sample standard deviation).[11] The results confirm that higher social media sentiment reduces informational efficiency, deviating the prices from the fundamental values and lowering the information incorporation process.

[Insert Table 3 Here]

4.2. Social media sentiment constructed using alternative dictionaries

We first assess the robustness of our main results to the choice of the natural language processing dictionary. Different dictionaries may differ in the way they extract the tone from a text (Bukovina, 2016). This can influence the measurement of social media sentiment, and accordingly, its predictive power. We extract the tone score of tweets using the three following dictionaries: Harvard IV-4 dictionary (Tetlock, 2007), Loughran-McDonald sentiment list (Loughran and McDonald, 2014), and SentiWordNet (Azar and Lo, 2016). After each tweet is classified into positive, negative or neutral categories via each new method, we aggregate them to a daily level following Equation (1), respectively.

Table 4 reports the regression results for autocorrelation (Panel A) and variance ratio (Panel B) on social media sentiment indices constructed using three different dictionaries. Our

[11] This is calculated as $0.015 \times 0.31 = 0.005$ where the regression coefficient (0.015) is multiplied by the standard deviation of the sentiment index (0.31) reported in Table 1. This is equivalent to a $(0.015 \times 0.31) \div 0.09 = 5.2\%$ increase in variance ratio, where 0.09 is the full sample standard deviation of variance ratio shown in Table 1.

results are robust to the choice of the natural language processing dictionary. More specifically, all three new sentiment indices have a positive impact on both informational efficiency measures. They are also statistically significant in most cases. These results provide support that a higher sentiment is associated with a lower informational efficiency.

[Insert Table 4 Here]

4.3. Social media sentiment constructed using different time period windows

In our main specification, we construct sentiment using tweets posted during the trading hours between 9:30 to 16:00 EST. The choice of this time interval may affect the degree of social media sentiment and therefore, our findings. To alleviate this concern, we reconstruct the daily social media sentiment index using tweets posted during different intraday time intervals. First, we consider tweets posted during the previous day from 00:00 to 23:59:59 EST. Second, we consider tweets posted during the pre-market period from 00:00 to 09:30 EST, i.e., tweets posted just before the market opens and the market information measures are calculated. Similar to the previous, each tweet is classified into positive, negative or neutral categories. We aggregate them to a daily level following Equation (1) and estimate Equation (4) for the two market efficiency measures with the newly constructed sentiment indices. The results are reported in Table 5.

[Insert Table 5 Here]

Our findings remain robust regardless of the periods used to construct social media sentiment. Sentiment positively and significantly affects autocorrelation (Panel A) and variance ratio (Panel B), both during the full day and the pre-market open. However, we acknowledge that the effect is weaker for the latter. This is likely due to lower Twitter activity before 9:30AM (see Figure 1.B.). However, the finding that tweets from pre-market open impacts autocorrelation implies that although there is less Twitter activity before the market opens, this information is still useful in explaining the same-day informational efficiency.

Overall, our findings remain robust regardless of the language dictionary used to extract social media sentiment, and the time period window used to construct the social media sentiment. Therefore, we conclude that a higher social media sentiment reduces informational efficiency.

5. Social Media Sentiment and Investor Herding

In this section, we explore the mechanism underlying the linkages between social media sentiment and informational efficiency. Previous studies document that psychological and social forces may explain aggregate financial market behavior (see, e.g., Fenzl and Pelzmann 2012; Filip and Pochea, 2023). Fenzl and Pelzmann (2012), for instance, demonstrate that nonmean-reverting dynamism in financial markets can result from irrational herding impulses sensed by market participants in complex and uncertain situations. Filip and Pochea (2023) further show that herding is a persistent phenomenon in the U.S. and European stock markets. Herding behavior occurs under both extreme positive and negative sentiments. Based on this evidence, we argue that high social media sentiment may accelerate herding behavior. Subsequently, it will cause prices to deviate from fundamental values and lower informational efficiency.

To investigate whether investors' herding is the mechanism that explains the negative relationship between social media sentiment and market efficiency, we employ the following vector autoregressive (VAR) model, (see Kurov, 2008; and Blasco et al., 2012)[12],

[12] Kurov (2008) uses a VAR model to capture the relationship between net order flows and returns, finding significant evidence of feedback trading. Blasco et al. (2012) uses a VAR model to explore the herding-sentiment connection and herding-return relation and find that sentiment and past returns drive herding behaviors among investors.

$$Herding_t = \zeta_1 + \sum_{i=1}^{l} \lambda_i Sentiment_{t-i} + \sum_{i=1}^{l} \mu_i Herding_{t-i} + \varepsilon_{1,t},$$

$$Sentiment_t = \zeta_2 + \sum_{i=1}^{l} \eta_i Sentiment_{t-i} + \sum_{i=1}^{l} \theta_i Herding_{t-i} + \varepsilon_{2,t}, \tag{5}$$

where $Herding_t$ is one of the herding measures on day t. We use five lags based on the Schwartz Information Criterion (SIC).

There are several measures for capturing herding activity. We construct two different herding indicators following Cai et al. (2019) and Zhou (2018).[13] Inspired by Lakonishok et al. (1992), Cai et al. (2019) develop the dollar-based herding (*DH*) measure. This measure considers trading volume for measuring the intensity of herding behavior and is measured as follows,

$$DH_t = \frac{|Buy\ Amount_t - Sell\ Amount_t|}{Buy\ Amount_t + Sell\ Amount_t}, \tag{6}$$

where DH_t is the herding on day t, measured as the absolute difference between buyer-initiated ($Buy\ Amount_t$) and seller-initiated ($Sell\ Amount_t$) dollar volumes. A higher DH value indicates higher degree of herding intensity.

Second, we use the Williams Percent Range (WR) which measures herding activity as follows,

$$WR_t = -\frac{P_{t-1,t-11}^{high} - p_t^{close}}{P_{t-1,t-11}^{high} - P_{t-1,t-11}^{low}} \times 100 \tag{7}$$

[13] We only consider one asset (SPY) and, therefore, we cannot calculate herding measures such as cross-sectional absolute deviations and cross-sectional standard deviations which require a cross section of assets (see e.g., Christie and Huang, 1995; Chang et al., 2000; or Lakonishok et al., 1992).

where the $P_{t-1,t-11}^{high}$ and $P_{t-1,t-11}^{low}$ are the highest and lowest prices over the prior ten days, from $t-11$ to $t-1$, and p_t^{close} is the closing price on day t. To ease interpretation, we multiply this metric by -1. Thus, a higher (lower) WR represents a higher intensity of overbought (oversold). Following Zhou (2018), if WR is greater than -20, the asset is regarded as overbought, and when WR is less than -80, the asset is regarded as oversold. We report the VAR results from Equation (5) in Panels A and B of Table 6, respectively.

[Insert Table 6 Here]

Turning first to Panel A, we observe that the coefficient of the first sentiment lag is significant and positive at 0.059 (t-statistic of 2.40) for the DH_t and 0.032 (t-statistic of 2.15) for WR_t. This indicates that a higher sentiment is associated with a higher herding behaviour (DH) and greater likelihood of overbought (WR) the following day. This result implies that investors mimic the trading of others during optimistic periods. This collective reaction eventually leads to an imbalance between buy and sell transactions, causing prices to deviate from their fundamental values. We interpret the positive effect of sentiment on both herding measures being caused by short-selling constraints. In particular, while investors may herd during optimistic times, they are unable to short sell during pessimistic times. This is consistent with Barber and Odean (2008) who explain that individual investors tend to be net buyers for stocks experiencing high abnormal trading volume and stocks with extreme returns, but they can only sell stocks they already own. We do not observe any reserve causality from herding to sentiment. This observation is consistent with Blasco et al. (2012) who find that investor sentiment Granger causes herding but not vice versa.

To understand the dynamic relationship between sentiment and herding behaviour, we follow the literature and plot the cumulative generalized impulse response functions (Pesaran and Shin, 1998). Specifically, we plot how one standard deviation shock in the social media sentiment impacts the DH and WR herding measures in Figures 3 and 4, respectively.

[Insert Figure 3 Here]

[Insert Figure 4 Here]

Figure 3 illustrates that a one standard deviation shock from social media sentiment significantly increases the dollar-based herding in the following 8 days before the effect disappears. We find a similar pattern with the WR herding measure in Figure 4, but this effect only lasts for two days. These results suggest that an optimistic view on the SPY leads to a significant and short-lived increase in herding behaviour among investors. As a result, informational efficiency decreases.

One potential concern about the WR herding measure is that we treat WR as continuous variables in the VAR model specification in Equation (5). Zhou (2018) argues that WR score higher than -20 is defined as overbought and a score lower than -80 is classified as oversold. Given that an overbought and an oversold asset can imply investors' herding behavior, we redefine the WR herding measure as an indicator variable which equals one on days when the WR value is less than -80 or greater than -20, and zero otherwise. We run a logistic regression using the WR herding indicator as the dependent variable, and the lagged social media sentiment as the main independent variable. We employ the same control variables as with Equation (4).

We report the coefficients, odds ratio and t-statistics of the logistic regression in Table 7. Odds ratio is the exponentiated coefficient, representing the proportional change of parameters. The coefficient for the lagged sentiment is positive at 0.045 (with an odds ratio of 1.046), statistically significant at the 1% level. In probabilistic terms, this coefficient can be interpreted as a one unit increase in social media sentiment is associated with a higher likelihood of herding behavior the following day by a factor of 1.046. In other words, the probability of herding increases by a factor of 74% with an increase of one standard deviation in sentiment[14]. Furthermore, the Pseudo-R^2 is high at 0.80, indicating this model fits the data well. These results are in line with the findings of the VAR model reported in Table 6.

[Insert Table 7 Here]

In summary, we demonstrate that a higher social media sentiment reduces informational efficiency the following day. This relationship can be explained by the increase in investor herding behavior which contributes to a decline in the quality of the information environment, resulting in informationally inefficient prices.

6. Robustness tests

In this section, we conduct multiple robustness tests on our main results, considering market conditions, the asymmetry effect in the social media effect on informational efficiency.

6.1. Social media sentiment and market efficiency with considering market controls

we have demonstrated that higher social media sentiment increases market informational inefficiency on the following day, as reported in Table 2 and Table 3. However, there is a possibility that both social media sentiment and market efficiency are driven by the state of the business cycles and market conditions, which may affect our findings. To ensure that our results are not driven by market fundamentals, we control for the term spread

[14] We use sigmoid function to calculate and obtain 74%. Specifically, it equals $\frac{1}{1+e^{-1.046}}$, where e is Euler's number.

(calculated as the difference between the 10-year and 2-year Treasure Bonds), the 3-month Treasury Bill, and the TED spread which is a proxy of funding liquidity. All these business cycle variables are downloaded from the Federal Reserve Bank of St. Louis.

Table 8 and Table 9 report the regression estimates of Equation (4) after controlling for the above three market factors, with the autocorrelation and variance ratio as the dependent variable, respectively. The results suggest that our main result holds after controlling for the state of the business cycle and market conditions, confirming the impact of social media sentiment on market informational inefficiency.

[Insert Table 8 Here]

[Insert Table 9 Here]

6.2. The asymmetric effect of positive and negative social media sentiment on market efficiency

We also investigate whether the impact of social media sentiment on market informational efficiency is related to the characteristic of sentiment, specifically the asymmetric effect of positive and negative social media sentiment. Existing literature has documented the debate on the significance of positive and negative sentiment on stock market impact. On one hand, positive sentiment could have a stronger effect than negative sentiment due to retail investors' limited cognitive processing ability (Barber and Odean, 2008). On the other hand, negative sentiment is also pronounced and dominant due to the heuristics of investors (Agrawal et al., 2018).

We decompose social media sentiment into positive and negative by classifying all tweets with positive and negative tone scores during the interval as $M_t^{Positive}$ and $M_t^{Negative}$, respectively. We then adjust Equation (4) and regress positive and negative sentiments with market efficiency measures as follows,

$$Y_t = \alpha + \beta \cdot Positive_{t-1} + \mu \cdot Negative_{t-1} + \delta \cdot Y_{t-1} + \gamma \cdot Controls_t + \varepsilon_t, \tag{8}$$

where Y_t is one of the two measures of informational efficiency on day t, i.e., $Autocorrelation^{PCA}$ or $VarianceRatio^{PCA}$. $Positive_{t-1}$ and $Negative_{t-1}$ are the decomposed sentiments $M_t^{Positive}$ and $M_t^{Negative}$ in natural logarithm at day $t-1$ and other control variables are the same as in Equation (4).

Table 10 reports the regression coefficients from Equation (8). The coefficients of positive sentiment are positive and significant at the 5% level for autocorrelation and variance ratio (0.020 and 0.013, respectively), while coefficients of negative sentiment are negative and significant at the 1% and 10% level for autocorrelation and variance ratio (-0.034 and -0.012, respectively). This finding indicates that both positive and negative sentiments affect market informational efficiency. However, positive sentiment increases market informational inefficiency and negative sentiment mitigates market informational inefficiency the following day. This can be due to the short-sales constraint among irrational investors, as they overreact to positive sentiment shocks but underreact to negative sentiment shocks (Baker and Stein, 2004). This reduces collective irrational trading in the market under negative sentiment and facilitates market informational efficiency. This confirms that social media sentiment impacts market informational efficiency, with an asymmetric effect between positive and negative sentiment.

[Insert Table 10 Here]

7. Conclusions

We examine the impact of social media sentiment on informational efficiency. We employ a natural language processing analysis to extract sentiment from tweets and analyze its impact on the efficiency of the SPDR S&P 500 ETF (SPY) prices. Using autocorrelation and variance ratio as informational efficiency measures, our findings indicate that higher social media sentiment reduces informational efficiency the following day. This finding highlights the influence of optimistic social media sentiment on market inefficiency.

We also delve into the underlying transmission channel. Our study shows that higher social media sentiment intensifies investors' herding activity the following day. Heightened sentiment leads to collective trading behaviors which result in one-sided buying or selling actions. Such herding behavior acts as an obstacle to the efficient dissemination of information and diminishes informational efficiency.

Our study has important implications for various stakeholders. For market participants, our findings highlight the importance of incorporating social media sentiment as a crucial factor when devising investment strategies. For regulators and policymakers, our study highlights the potential of social media as an additional surveillance tool within the market regulatory framework. Recognizing the influence of social media sentiment can aid in enhancing market oversight to effectively monitor and manage potential risks and market disruptions.

Reference list

AlMousa, M., Benlamri, R., & Khoury, R. (2021). Exploiting non-taxonomic relations for measuring semantic similarity and relatedness in WordNet. *Knowledge-Based Systems*, 212, 106565.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance*, 59(3), 1259-1294.

Azar, P.D. and Lo, A.W., 2016. The wisdom of Twitter crowds: predicting stock market reactions to FOMC meetings via Twitter feeds. *Journal of Portfolio Management*, 42(5), 123-134.

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *Journal of Finance*, 61(4), 1645-1680.

Baker, M., Wurgler, J., & Yuan, Y. (2012). Global, local, and contagious investor sentiment. *Journal of Financial Economics*, 104(2), 272-287.

Barber, B. M., & Odean, T. (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The review of financial studies*, 21(2), 785-818.

Barber, B. M., Odean, T., & Zhu, N. (2008). Do retail trades move markets? *Review of Financial Studies*, *22*(1), 151-186.

Baker, M., & Stein, J. C. (2004). Market liquidity as a sentiment indicator. *Journal of financial Markets*, 7(3), 271-299.

Blasco, N., Corredor, P., & Ferreruela, S. (2012). Market sentiment: a key factor of investors' imitative behaviour. *Accounting & Finance*, 52(3), 663-689.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1-8.

Brogaard, J., Hagströmer, B., Nordén, L., & Riordan, R. (2015). Trading fast and slow: Colocation and liquidity. *Review of Financial Studies*, *28*(12), 3407-3443.

Bukovina, J. (2016). Social media big data and capital markets—An overview. *Journal of Behavioral and Experimental Finance*, 11, 18-26.

Busse, J. A., & Green, T. C. (2002). Market efficiency in real time. *Journal of Financial Economics*, 65(3), 415-437.

Cai, F., Han, S., Li, D., & Li, Y. (2019). Institutional herding and its price impact: Evidence from the corporate bond market. *Journal of Financial Economics*, 131(1), 139-167.

Chang, E. C., Cheng, J. W., & Khorana, A. (2000). An examination of herd behavior in equity markets: An international perspective. *Journal of Banking & Finance*, 24(10), 1651-1679.

Chau, F., Deesomsak, R., & Lau, M. C. (2011). Investor sentiment and feedback trading: Evidence from the exchange-traded fund markets. *International Review of Financial Analysis*, 20(5), 292-305.

Chen, H., De, P., Hu, Y., & Hwang, B. H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, 27(5), 1367-1403.

Christie, W. G., & Huang, R. D. (1995). Following the pied piper: do individual returns herd around the market? *Financial Analysts Journal*, 51(4), 31-37.

Chordia, T., Roll, R. and Subrahmanyam, A., 2001. Market liquidity and trading activity. *Journal of Finance*, 56(2), 501-530.

Clement, M. B., & Tse, S. Y. (2005). Financial analyst characteristics and herding behavior in forecasting. *Journal of Finance*, 60(1), 307-341.

Comerton-Forde, C., & Putniņš, T. J. (2015). Dark trading and price discovery. *Journal of Financial Economics*, 118(1), 70-92.

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *Journal of Finance*, 66(5), 1461-1499.

De Bondt, W. F., & Thaler, R. (1985). Does the stock market overreact? *Journal of Finance*, 40(3), 793-805.

De Jong, P., Elfayoumy, S., & Schnusenberg, O. (2017). From returns to tweets and back: An investigation of the stocks in the Dow Jones industrial average. *Journal of Behavioral Finance*, 18(1), 54-64.

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Positive feedback investment strategies and destabilizing rational speculation. *Journal of Finance*, 45(2), 379-395.

Duz Tan, S., & Tas, O. (2021). Social media sentiment in international stock returns and trading activity. *Journal of Behavioral Finance*, 22(2), 221-234.

Edmans, A., Garcia, D., & Norli, Ø. (2007). Sports sentiment and stock returns. *Journal of Finance*, 62(4), 1967-1998.

Edmans, A., Fernandez-Perez, A., Garel, A., & Indriawan, I. (2022). Music sentiment and stock returns around the world. *Journal of Financial Economics*, 145(2), 234-254.

Fenzl, T., & Pelzmann, L. (2012). Psychological and social forces behind aggregate financial market behavior. *Journal of Behavioral Finance*, 13(1), 56-65.

Filip, A. M., & Pochea, M. M. (2023). Intentional and spurious herding behavior: A sentiment driven analysis. *Journal of Behavioral and Experimental Finance*, 100810.

Foucault, T., Hombert, J., & Roşu, I. (2016). News trading and speed. *The Journal of Finance*, 71(1), 335-382.

Frijns, B., Indriawan, I., Tourani-Rad, A., & Zhang, H. (2023). The effect of equity market uncertainty on informational efficiency: Cross-sectional evidence. *Global Finance Journal*, 100854.

Gan, B., Alexeev, V., Bird, R., & Yeung, D. (2020). Sensitivity to sentiment: News vs social media. *International Review of Financial Analysis*, 67, 101390.

Grinblatt, M., Keloharju, M., & Linnainmaa, J. T. (2012). IQ, trading behavior, and performance. *Journal of Financial Economics*, 104(2), 339-362.

Grossman, S. J. (1995). Dynamic asset allocation and the informational efficiency of markets. *Journal of Finance*, 50(3), 773-787.

Gu, C., & Kurov, A. (2020). Informational role of social media: Evidence from Twitter sentiment. *Journal of Banking & Finance*, 121, 105969.

Guégan, D., & Renault, T. (2021). Does investor sentiment on social media provide robust information for Bitcoin returns predictability? *Finance Research Letters*, 38, 101494.

Hasbrouck, J. (1991). Measuring the information content of stock trades. *Journal of Finance*, 46(1), 179-207.

Hendershott, T., & Jones, C. M. (2005). Island goes dark: Transparency, fragmentation, and regulation. *The Review of Financial Studies*, 18(3), 743-793.

Hendershott, T., Jones, C. M., & Menkveld, A. J. (2011). Does algorithmic trading improve liquidity? *Journal of Finance*, *66*(1), 1-33.

Hirshleifer, D., & Shumway, T. (2003). Good day sunshine: Stock returns and the weather. *Journal of Finance*, 58(3), 1009-1032.

Jiao, P., Veiga, A., & Walther, A. (2020). Social media, news media and the stock market. *Journal of Economic Behavior & Organization*, 176, 63-90.

Kaplanski, G., & Levy, H. (2014). Sentiment, irrationality and market efficiency: The case of the 2010 FIFA World Cup. *Journal of Behavioral and Experimental Economics*, 49, 35-43.

Karampatsas, N., Malekpour, S., Mason, A., & Mavis, C. P. (2023). Twitter investor sentiment and corporate earnings announcements. *European Financial Management*, 29(3), 953-986.

Kim, J. S., Ryu, D., & Seo, S. W. (2014). Investor sentiment and return predictability of disagreement. *Journal of Banking & Finance*, 42, 166-178.

Kumar, A., & Lee, C. M. (2006). Retail investor sentiment and return comovements. *Journal of Finance*, 61(5), 2451-2486.

Kurov, A. (2008). Investor sentiment, trading behavior and informational efficiency in index futures markets. *Financial Review*, 43(1), 107-127.

Lakonishok, J., Shleifer, A., & Vishny, R. W. (1992). The impact of institutional trading on stock prices. *Journal of Financial Economics*, 32(1), 23-43.

Lee, C. M., & Ready, M. J. (1991). Inferring trade direction from intraday data. *The Journal of Finance*, 46(2), 733-746.

Leung, H., & Ton, T. (2015). The impact of internet stock message boards on cross-sectional returns of small-capitalization stocks. *Journal of Banking & Finance*, 55, 37-55.

Li, T., Chen, H., Liu, W., Yu, G., & Yu, Y. (2023). Understanding the role of social media sentiment in identifying irrational herding behavior in the stock market. *International Review of Economics and Finance*, 87, 163-179.

Lo, A., & MacKinlay, C. (1988). Stock market prices do not follow random walks: evidence from a simple specification test. *Review of Financial Studies*, 1,41–66.

Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. *Journal of Finance*, 69(4), 1643-1671.

McKenzie, M. D., & Faff, R. W. (2003). The determinants of conditional autocorrelation in stock returns. *Journal of Financial Research*, 26(2), 259-274.

McKenzie, M. D., & Kim, S. J. (2007). Evidence of an asymmetry in the relationship between volatility and autocorrelation. *International Review of Financial Analysis*, 16(1), 22-40.

Menkhoff, L., & Nikiforow, M. (2009). Professionals' endorsement of behavioral finance: does it impact their perception of markets and themselves? *Journal of Economic Behavior & Organization*, 71(2), 318-329.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2), 1-69.

Nofsinger, J. R., & Sias, R. W. (1999). Herding and feedback trading by institutional and individual investors. *Journal of Finance*, 54(6), 2263-2295.

Nofsinger, J. R. (2005). Social mood and financial economics. *Journal of Behavioral Finance*, 6(3), 144-160.

O'Hara, M., & Ye, M. (2011). Is market fragmentation harming market quality?. *Journal of Financial Economics*, 100(3), 459-474.

Pesaran, H.H. and Shin, Y., 1998. Generalized impulse response analysis in linear multivariate models. *Economics Letters*, 58, 17-29.

Roll, R. (1984). A simple implicit measure of the effective bid-ask in an efficient market. *Journal of Finance*, 39,1127–1139.

Shen, J., Yu, J., & Zhao, S. (2017). Investor sentiment and economic forces. *Journal of Monetary Economics*, 86, 1-21.

Shiller, R. J. (1981). Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends? *The American Economic Review*, 71, 421–436.

Siganos, A., Vagenas-Nanos, E. and Verwijmeren, P., 2014. Facebook's daily sentiment and international stock markets. *Journal of Economic Behavior & Organization*, 107, 730-743.

Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management*, 20(5), 926-957.

Stambaugh, R. F., Yu, J., & Yuan, Y. (2012). The short of it: Investor sentiment and anomalies. *Journal of Financial Economics*, 104(2), 288-302.

Sun, L., Najand, M., & Shen, J. (2016). Stock return predictability and investor sentiment: A high-frequency perspective. *Journal of Banking & Finance*, 73, 147-164.

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

Tetlock, P. C. (2011). All the news that's fit to reprint: Do investors react to stale information?. *The Review of Financial Studies*, 24(5), 1481-1512.

Valadkhani, A. (2022). Do large-cap exchange-traded funds perform better than their small-cap counterparts in extreme market conditions? *Global Finance Journal*, 53, 100743.

Vidhu Bhala, R. V., & Abirami, S. (2014). Trends in word sense disambiguation. *Artificial Intelligence Review*, 42, 159-171.

Vozlyublennaia, N. (2014). Investor attention, index performance, and return predictability. *Journal of Banking & Finance*, 41, 17-35.

Welch, I. (2000). Herding among security analysts. *Journal of Financial Economics*, 58(3), 369-396.

Zhang, X. F. (2006). Information uncertainty and stock returns. *Journal of Finance*, 61(1), 105-137.

Zhou, G. (2018). Measuring investor sentiment. *Annual Review of Financial Economics*, 10, 239-259.

Table 1. Summary statistics and correlation table

This table reports the daily summary statistics for tweets and market variables across the sample period from August 1, 2012, to March 31, 2022. $Sentiment$ is the social media sentiment index from Equation (1). $Autocorrelation^{PCA}$ is the first principal component of absolute midquote return autocorrelation constructed using various frequencies, e.g., 1-, 10-, and 30-sec, for each trading day from 9:40 to 15:50. $VarianceRatio^{PCA}$ is the first principal component of variance ratio constructed using various frequencies, e.g., [10-sec, 30-sec], [30-sec, 60-sec] and [10-sec, 60-sec] for each trading day from 9:40 to 15:50 EST. Both metrics are scaled so that they range from zero to one. $Return$ is the daily return of SPY, $Volatility$ is the realized volatility constructed using SPY midquote returns at a one-minute frequency, $Volume$ is the log of daily total dollar volume, $Depth$ is the log of daily average bid-ask depth, VIX is the daily S&P 500 implied volatility index.

	$Sentiment$	$Autocorrelation^{PCA}$	$VarianceRatio^{PCA}$	$Return$	$Volatility$	$Volume$	$Depth$	VIX
Panel A Descriptive statistics								
Mean	0.76	0.16	0.12	0.00	0.01	23.67	8.24	17.18
Std. dev.	0.31	0.13	0.09	0.01	0.01	0.43	0.70	7.06
Median	0.76	0.13	0.10	0.00	0.01	23.61	8.13	15.20
5th Percentile	0.28	0.02	0.01	-0.02	0.00	23.07	7.26	10.63
95th Percentile	1.23	0.39	0.29	0.01	0.01	24.45	9.57	29.32
AR(1)	0.39	0.06	0.06	-0.14	0.46	0.71	0.91	0.97
Obs.	2,432	2,432	2,432	2,432	2,432	2,432	2,432	2,432
Panel B Correlation Matrix								
$Sentiment$	1							
$Autocorrelation^{PCA}$	0.04	1						
$VarianceRatio^{PCA}$	0.03	0.27	1					
$Return$	0.23	0.03	0.03	1				
$Volatility$	-0.05	-0.02	-0.02	-0.06	1			
$Volume$	-0.03	-0.07	-0.02	-0.19	0.22	1		
$Depth$	0.14	-0.10	-0.03	-0.13	0.17	0.72	1	
VIX	0.02	-0.06	-0.04	-0.17	0.26	0.62	0.69	1

Table 2. Autocorrelation and social media sentiment

This table reports the coefficient estimates of Equation (4) with the first principal component of autocorrelation, $Autocorrelation_t^{PCA}$ as the market efficiency measure. It is constructed using 1-, 10-, and 30-sec absolute midquote autocorrelations for each trading day from 9:40 to 15:50 EST. The metric is scaled so that it ranges from zero to one. $Sentiment$ is the social media sentiment index from Equation (1). $Return$ is the daily return of SPY, $Volatility$ is the realized volatility constructed using SPY midquote returns at a one-minute frequency, $Volume$ is the log of daily total dollar volume, $Depth$ is the log of daily average bid-ask depth, VIX is the daily S&P 500 implied volatility index. The Newey-West corrected t-statistics are in parenthesis. ***, ** and * indicate 1%, 5% and 10% significance level.

	Dependent: $Autocorrelation_t^{PCA}$													
	(1)		(2)		(3)		(4)		(5)		(6)		(7)	
$Sentiment_{t-1}$	0.022**	(2.26)	0.021**	(2.26)	0.019**	(2.05)	0.021**	(2.33)	0.028***	(3.08)	0.022**	(2.44)	0.030***	(3.21)
$Autocorrelation_{t-1}^{PCA}$			0.061**	(2.36)	0.057**	(2.25)	0.054**	(2.20)	0.051**	(2.11)	0.057**	(2.23)	0.052**	(2.19)
$Return_t$			0.055*	(1.88)									0.393	(1.33)
$Volatilty_t$					-1.044**	(-2.32)							0.897	(0.83)
$Volume_t$							-0.021***	(-3.35)					-0.003	(-0.27)
$Depth_t$									-0.018***	(-4.44)			-0.021***	(-2.96)
VIX_t											-0.001***	(-2.91)	0.001	(0.28)
Adj. R^2	0.002		0.006		0.007		0.007		0.010		0.014		0.014	
Obs.	2,432		2,432		2,432		2,432		2,432		2,432		2,432	

Table 3. Variance Ratio and social media sentiment

This table reports the coefficient estimates of Equation (4) with the first principal component of variance ratio, $VarianceRatio_t^{PCA}$ as the market efficiency measure. It is constructed using the absolute variance ratio of [10-sec, 30-sec], [30-sec, 60-sec] and [10-sec, 60-sec] for each trading day from 9:40 to 15:50 EST and the metric is scaled from zero to one. $Sentiment$ is the social media sentiment index from Equation (1). $Return$ is the daily return of SPY, $Volatility$ is the realized volatility constructed using SPY midquote returns at a one-minute frequency, $Volume$ is the log of daily total dollar volume, $Depth$ is the log of daily average bid-ask depth, VIX is the daily S&P 500 implied volatility index. The Newey-West corrected t-statistics are in parenthesis. ***, ** and * indicate 1%, 5% and 10% significance level.

	Dependent: $VarianceRatio_t^{PCA}$													
	(1)		(2)		(3)		(4)		(5)		(6)		(7)	
$Sentiment_{t-1}$	0.013**	(2.00)	0.012*	(1.95)	0.011*	(1.81)	0.012**	(1.96)	0.015**	(2.28)	0.013**	(2.07)	0.015**	(2.27)
$VarianceRatio_{t-1}^{PCA}$			0.053***	(2.85)	0.051***	(2.80)	0.052***	(2.75)	0.051***	(2.79)	0.051***	(2.62)	0.052***	(2.77)
$Return_t$			0.410*	(1.83)									0.377	(1.61)
$Volatilty_t$					-0.066	(-1.42)							0.468	(0.56)
$Volume_t$							-0.005	(-1.18)					0.004	(0.55)
$Depth_t$									-0.006**	(-2.17)			-0.004	(-0.76)
VIX_t											-0.001**	(-2.51)	-0.001	(-1.17)
Adj. R^2	0.001		0.005		0.004		0.004		0.005		0.006		0.006	
Obs.	2,432		2,432		2,432		2,432		2,432		2,432		2,432	

Table 4. Informational efficiency and social media sentiment constructed using alternative dictionaries

This table reports the coefficient estimates of Equation (4) with two market efficiency measures, $Autocorrelation_t^{PCA}$ (Panel A) and $VarianceRatio_t^{PCA}$ (Panel B). Both metrics are constructed for each trading day from 9:40 to 15:50 EST and scaled from zero to one. $Sentiment$ is the social media sentiment index from Equation (1) constructed using one of the three different dictionaries, the Harvard IV-4 sentiment list (Tetlock, 2007), the SentiWordNet (Azar and Lo, 2016), and the Loughran-McDonald sentiment list (Loughran and McDonald, 2014). $Return$ is the daily return of SPY, $Volatility$ is the realized volatility constructed using SPY midquote returns at a one-minute frequency, $Volume$ is the log of daily total dollar volume, $Depth$ is the log of daily average bid-ask depth, VIX is the daily S&P 500 implied volatility index. The Newey-West corrected t-statistics are in parenthesis. ***, ** and * indicate 1%, 5% and 10% significance level.

	Panel A: $Autocorrelation_t^{PCA}$						Panel B: $VarianceRatio_t^{PCA}$					
	Harvard IV-4		*SentiWordNet*		*Loughran-McDonald*		*Harvard IV-4*		*SentiWordNet*		*Loughran-McDonald*	
$Sentiment_{t-1}$	0.022**	(2.48)	0.021*	(1.74)	0.024***	(3.20)	0.015**	(2.13)	0.016*	(1.89)	0.007	(1.29)
$Dependent_{t-1}$	0.054**	(2.32)	0.054**	(2.27)	0.054**	(2.32)	0.053***	(2.85)	0.052***	(2.76)	0.053***	(2.82)
$Return_t$	0.358	(1.19)	0.377	(1.26)	0.386	(1.29)	0.357	(1.53)	0.368	(1.60)	0.372	(1.59)
$Volatilty_t$	0.660	(0.60)	0.415	(0.38)	0.720	(0.67)	0.417	(0.50)	0.272	(0.33)	0.279	(0.34)
$Volume_t$	-0.002	(-0.16)	-0.004	(-0.36)	-0.003	(-0.32)	0.005	(0.68)	0.004	(0.54)	0.004	(0.48)
$Depth_t$	-0.022***	(-3.01)	-0.019***	(-2.74)	-0.021***	(-3.03)	-0.006	(-1.09)	-0.004	(-0.79)	-0.003	(-0.61)
VIX_t	0.001	(0.30)	0.001	(0.51)	0.001	(0.42)	-0.001	(-1.18)	-0.001	(-1.00)	-0.001	(-1.08)
Adj. R^2	0.012		0.011		0.014		0.006		0.005		0.004	
Obs.	2,432		2,432		2,432		2,432		2,432		2,432	

Table 5. Informational efficiency and social media sentiment constructed using alternative intervals

This table reports the coefficient estimates of Equation (4) with two market efficiency measures, $Autocorrelation_t^{PCA}$ (Panel A) and $VarianceRatio_t^{PCA}$ (Panel B). $Sentiment$ is the social media sentiment index from Equation (1) constructed using alternative intervals, from 00:00 to 23:59:59 EST (Full day) or from 00:00 to 09:29:59 EST (Pre-market open). $Return$ is the daily return of SPY, $Volatility$ is the realized volatility constructed using SPY midquote returns at a one-minute frequency, $Volume$ is the log of daily total dollar volume, $Depth$ is the log of daily average bid-ask depth, VIX is the daily S&P 500 implied volatility index. The Newey-West corrected t-statistics are in parenthesis. ***, ** and * indicate 1%, 5% and 10% significance level.

	Panel A: $Autocorrelation_t^{PCA}$				Panel B: $VarianceRatio_t^{PCA}$			
	Full day		Pre-market open		Full day		Pre-market open	
$Sentiment_{t-1}$	0.033***	(3.20)	0.014**	(2.29)	0.013*	(1.86)	0.006	(1.25)
$Dependent_{t-1}$	0.052**	(2.25)	0.054**	(2.29)	0.052***	(2.77)	0.053***	(2.84)
$Return_t$	0.400	(1.33)	0.296	(0.97)	0.378	(1.62)	0.335	(1.42)
$Volatilty_t$	1.064	(0.97)	0.68	(0.63)	0.454	(0.54)	0.322	(0.39)
$Volume_t$	-0.002	(-0.17)	-0.003	(-0.31)	0.005	(0.57)	0.004	(0.51)
$Depth_t$	-0.021***	(-2.98)	-0.019***	(-2.71)	-0.003	(-0.68)	-0.003	(-0.54)
VIX_t	0.001	(0.20)	0.001	(0.23)	-0.001	(-1.19)	-0.001	(-1.19)
Adj. R^2	0.014		0.012		0.005		0.004	
Obs.	2,432		2,432		2,432		2,432	

Table 6. Investor herding and social media sentiment

This table reports the coefficient estimates of Equation (5) with two herding measures, Dollar Based Herding, DH (Panel A) and Williams Percent Range, WR (Panel B), as described in Equation (6) and (7), respectively. Higher DH indicates a higher level of herding. Higher (lower) WR indicates the asset is overbought (oversold). $Sentiment$ is the social media sentiment index from Equation (1). The Schwartz Information Criterion (SIC) is used to choose the optimal number of lags. All variables are normalized. The Newey-West corrected t-statistics are in parenthesis. ***, ** and * indicate 1%, 5% and 10% significance level.

	Panel A: DH_t				Panel B: WR_t			
	$Herding_t$		$Sentiment_t$		$Herding_t$		$Sentiment_t$	
$Sentiment_{t-1}$	0.059^{**}	(2.40)	0.235^{***}	(9.48)	0.032^{**}	(2.15)	0.224^{***}	(8.13)
$Sentiment_{t-2}$	0.005	(0.23)	0.140^{***}	(6.24)	-0.02	(-1.28)	0.138^{***}	(5.80)
$Sentiment_{t-3}$	0.021	(0.94)	0.137^{***}	(6.17)	-0.008	(-0.53)	0.143^{***}	(6.10)
$Sentiment_{t-4}$	-0.043^{*}	(-1.74)	0.111^{***}	(4.89)	0.023	(1.37)	0.119^{***}	(4.96)
$Sentiment_{t-5}$	-0.012	(-0.47)	0.103^{***}	(4.67)	0.003	(0.22)	0.114^{***}	(4.99)
$Herding_{t-1}$	0.106^{***}	(5.27)	0.006	(0.32)	0.700^{***}	(31.37)	0.033	(1.13)
$Herding_{t-2}$	0.060^{***}	(2.84)	0.015	(0.84)	0.092^{***}	(3.23)	-0.005	(-0.15)
$Herding_{t-3}$	0.038	(1.62)	-0.033^{*}	(-1.74)	-0.003	(-0.13)	-0.034	(-0.93)
$Herding_{t-4}$	0.055^{***}	(2.64)	0.004	(0.22)	-0.028	(-1.17)	-0.009	(-0.31)
$Herding_{t-5}$	0.072^{***}	(3.01)	-0.007	(-0.41)	-0.013	(-0.57)	-0.027	(-0.98)
Adj. R^2	0.034		0.272		0.576		0.274	
Obs.	2432		2432		2432		2432	

Table 7. Logistic regression of investor herding and social media sentiment using classified Williams Percent Range

This table reports the coefficient estimates of logistic regression of categorized William Percent Range (WR) herding and social media sentiment. The estimate specification is similar to Equation (4) but with the dependent variable being the herding metric $Herding_t$, which takes a value of 1 if WR is less than -80 (i.e., oversold) or greater than -20 (overbought), and 0 otherwise. $Sentiment_{t-1}$ is the lagged social media sentiment index from Equation (1). $Return$ is the daily return of SPY, $Volatility$ is the realized volatility constructed using SPY midquote returns at a one-minute frequency, $Volume$ is the log of daily total dollar volume, $Depth$ is the log of daily average bid-ask depth, VIX is the daily S&P 500 implied volatility index. Odds ratios are the exponentiated coefficients, representing the proportional change of parameters. The Pseudo-R^2 measure ranges from 0 to 1, a higher value indicates a better fit of the model to the data. The Newey-West corrected t-statistics are in parenthesis. ***, ** and * indicate 1%, 5% and 10% significance level.

	Dependent: $Herding_t$		
	Coef.	Odds ratio	t-stat
$Sentiment_{t-1}$	0.045***	1.046	(4.57)
$Herding_{t-1}$	0.353***	1.423	(18.21)
$Return_t$	-0.012	0.988	(-1.08)
$Volatility_t$	-0.082***	0.921	(-4.14)
$Volume_t$	0.055***	1.056	(3.22)
$Depth_t$	-0.067***	0.935	(-3.96)
VIX_t	0.058***	1.060	(3.12)
Pseudo R^2		0.80	
Obs.		2432	

Table 8. Autocorrelation and social media sentiment with market controls

This table reports the coefficient estimates based on Equation (4) with controlling three market business cycle indicators. The first principal component of autocorrelation, $Autocorrelation_t^{PCA}$ is the market efficiency measure. It is constructed using 1-, 10-, and 30-sec absolute midquote autocorrelations for each trading day from 9:40 to 15:50 EST. The metric is scaled so that it ranges from zero to one. $Sentiment$ is the social media sentiment index from Equation (1). $Return$ is the daily return of SPY, $Volatility$ is the realized volatility constructed using SPY midquote returns at a one-minute frequency, $Volume$ is the log of daily total dollar volume, $Depth$ is the log of daily average bid-ask depth, VIX is the daily S&P 500 implied volatility index. The market controls, *Term spread* is calculated as the difference between the 10-year and 2-year Treasure Bonds, the *3M T-Bill* is 3-month Treasury Bill, and the *TED* is a proxy of funding liquidity. All these business cycle variables are from the Federal Reserve Bank of St. Louis. The Newey-West corrected t-statistics are in parenthesis. ***, ** and * indicate 1%, 5% and 10% significance level. *TED* was discontinued from 24 Jan 2022, resulting 109 less observations in regression (8) with controls.

	Dependent: $Autocorrelation_t^{PCA}$															
	(1)		(2)		(3)		(4)		(5)		(6)		(7)		(8)	
$Sentiment_{t-1}$	0.022**	(2.26)	0.021**	(2.26)	0.019**	(2.05)	0.021**	(2.33)	0.028***	(3.08)	0.022**	(2.44)	0.030***	(3.21)	0.034***	(3.49)
$Autocorrelation_{t-1}^{PCA}$			0.061**	(2.36)	0.057**	(2.25)	0.054**	(2.20)	0.051**	(2.11)	0.057**	(2.23)	0.052**	(2.19)	0.041*	(1.72)
$Return_t$			0.055*	(1.88)									0.393	(1.33)	0.385	(1.24)
$Volatilty_t$					-1.044**	(-2.32)							0.897	(0.83)	0.986	(0.77)
$Volume_t$							-0.021***	(-3.35)					-0.003	(-0.27)	0.004	(0.34)
$Depth_t$									-0.018***	(-4.44)			-0.021***	(-2.96)	-0.028***	(-3.09)
VIX_t											-0.001***	(-2.91)	0.001	(0.28)	0.000	(0.33)
Term spread															-0.003	(-0.36)
3M T-Bill															-0.01*	(-1.80)
TED															-0.013	(-0.55)
Adj. R^2	0.002		0.006		0.007		0.007		0.010		0.014		0.014		0.013	
Obs.	2,432		2,432		2,432		2,432		2,432		2,432		2,432		2324	

Table 9. Variance Ratio and social media sentiment with market controls

This table reports the coefficient estimates based on Equation (4) with controlling three market business cycle indicators. The first principal component of variance ratio, $VarianceRatio_t^{PCA}$ is the market efficiency measure. It is constructed using the absolute variance ratio of [10-sec, 30-sec], [30-sec, 60-sec] and [10-sec, 60-sec] for each trading day from 9:40 to 15:50 EST and the metric is scaled from zero to one. $Sentiment$ is the social media sentiment index from Equation (1). $Return$ is the daily return of SPY, $Volatility$ is the realized volatility constructed using SPY midquote returns at a one-minute frequency, $Volume$ is the log of daily total dollar volume, $Depth$ is the log of daily average bid-ask depth, VIX is the daily S&P 500 implied volatility index. The market controls, *Term spread* is calculated as the difference between the 10-year and 2-year Treasure Bonds, the *3M T-Bill* is 3-month Treasury Bill, and the TED is a proxy of funding liquidity. All these business cycle variables are from the Federal Reserve Bank of St. Louis. The Newey-West corrected t-statistics are in parenthesis. ***, ** and * indicate 1%, 5% and 10% significance level. TED was discontinued from 24 Jan 2022, resulting 111 less observations in regression (8) with controls.

	Dependent: $VarianceRatio_t^{PCA}$															
	(1)		(2)		(3)		(4)		(5)		(6)		(7)		(8)	
$Sentiment_{t-1}$	0.013**	(2.00)	0.012*	(1.95)	0.011*	(1.81)	0.012**	(1.96)	0.015**	(2.28)	0.013**	(2.07)	0.015**	(2.27)	0.014**	(2.07)
$Autocorrelation_{t-1}^{PCA}$			0.053***	(2.85)	0.051***	(2.80)	0.052***	(2.75)	0.051***	(2.79)	0.051***	(2.62)	0.052***	(2.77)	0.068***	(3.46)
$Return_t$			0.410*	(1.83)									0.377	(1.61)	0.413*	(1.73)
$Volatilty_t$					-0.066	(-1.42)							0.468	(0.56)	1.120	(1.13)
$Volume_t$							-0.005	(-1.18)					0.004	(0.55)	0.010	(1.12)
$Depth_t$									-0.006**	(-2.17)			-0.004	(-0.76)	-0.010*	(-1.66)
VIX_t											-0.001**	(-2.51)	-0.001	(-1.17)	-0.001	(-1.34)
Term spread															-0.010*	(-1.76)
3M T-Bill															-0.006	(-1.37)
TED															-0.036*	(-1.84)
Adj. R^2	0.001		0.005		0.004		0.004		0.005		0.006		0.006		0.009	
Obs.	2,432		2,432		2,432		2,432		2,432		2,432		2,432		2324	

Table 10. Positive and Negative sentiment on social media and market efficiency

This table reports the coefficient estimates of Equation (4) with two market efficiency measures, $Autocorrelation_t^{PCA}$ (Panel A) and $VarianceRatio_t^{PCA}$ (Panel B), and with sentiment decomposed in Positive and Negative, respectively. $Positive$ and $Negative$ are decomposed positive and negative social media sentiment indices similar to Equation (1) defined below with interval from 9:40 to 15:50 EST. $Return$ is the daily return of SPY, $Volatility$ is the realized volatility constructed using SPY midquote returns at a one-minute frequency, $Volume$ is the log of daily total dollar volume, $Depth$ is the log of daily average bid-ask depth, VIX is the daily S&P 500 implied volatility index. The Newey-West corrected t-statistics are in parenthesis. ***, ** and * indicate 1%, 5% and 10% significance level.

$$Positive_t = \ln\left[1 + M_t^{Positive}\right]$$

$$Negative_t = \ln\left[1 + M_t^{Negative}\right]$$

$$Y_t = \alpha + \beta \cdot Positive_{t-1} + \mu \cdot Negative_{t-1} + \delta \cdot Y_{t-1} + \gamma \cdot Controls_t + \varepsilon_t,$$

	Panel A: $Autocorrelation_t^{PCA}$		Panel B: $VarianceRatio_t^{PCA}$	
$Positive_{t-1}$	0.020**	(2.13)	0.013**	(1.97)
$Negative_{t-1}$	-0.034***	(-3.38)	-0.012*	(-1.72)
$Dependent_{t-1}$	0.051**	(2.24)	0.060***	(3.09)
$Return_t$	0.512	(1.60)	0.368	(1.54)
$Volatilty_t$	0.382	(0.34)	0.450	(0.53)
$Volume_t$	-0.001	(-0.08)	0.004	(0.53)
$Depth_t$	-0.014*	(-1.76)	-0.004	(-0.72)
VIX_t	0.001	(0.91)	-0.001	(-1.14)
Adj. R^2	0.014		0.005	
Obs.	2,432		2,432	

Figure 1. SPY tweets volume by different frequencies

Figure 1.A plots the number of SPY-related tweets by day of the week. Figure 1.B plots the SPY-related tweets by hour of the day. The sample period is from August 1, 2012, to March 31, 2022.

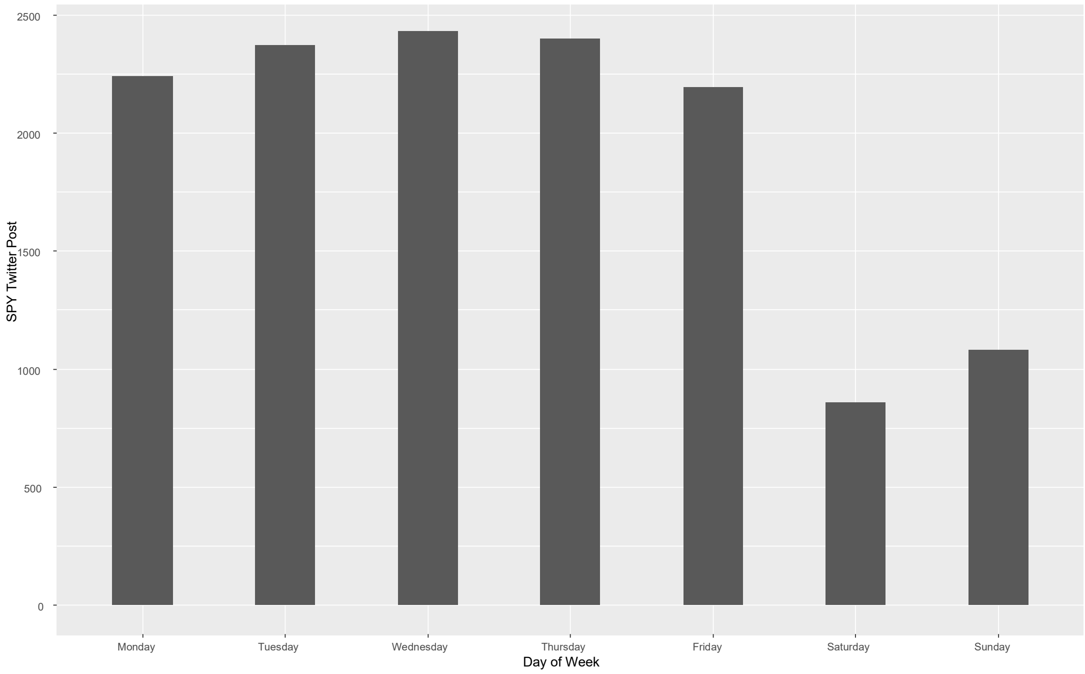

Figure 1.A. SPY-related tweets by day of the week

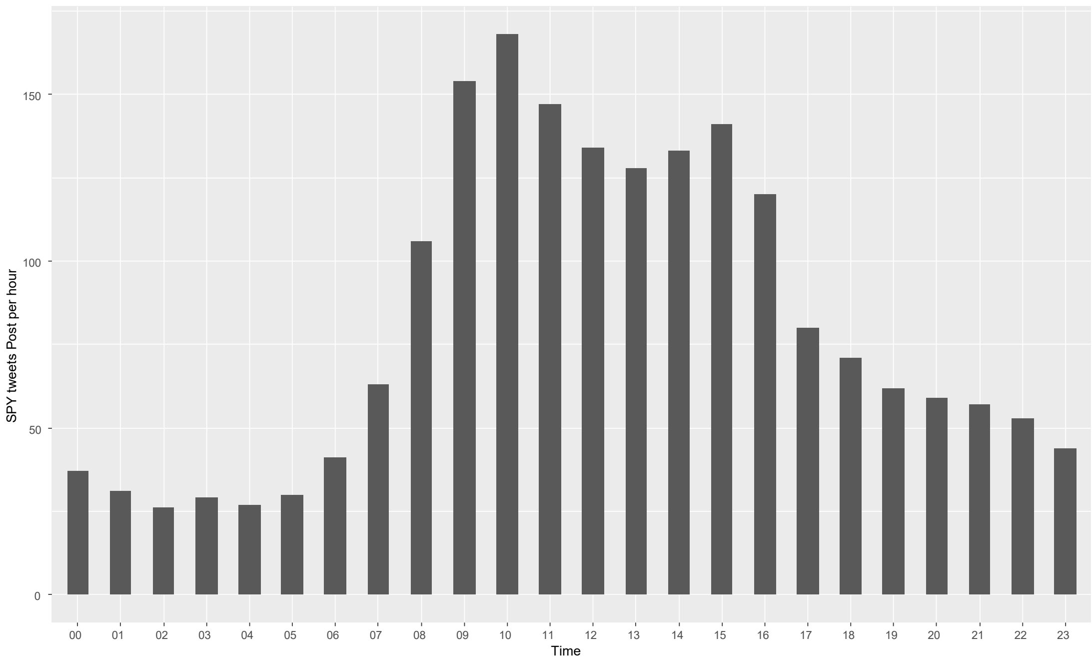

Figure 1.B. SPY-related tweets by the hour of the day

Figure 2. Social media sentiment and SPY Price

This figure plots the five-day moving average social media sentiment (dotted line) and daily SPY prices (solid line) from August 1, 2012, to March 31, 2022.

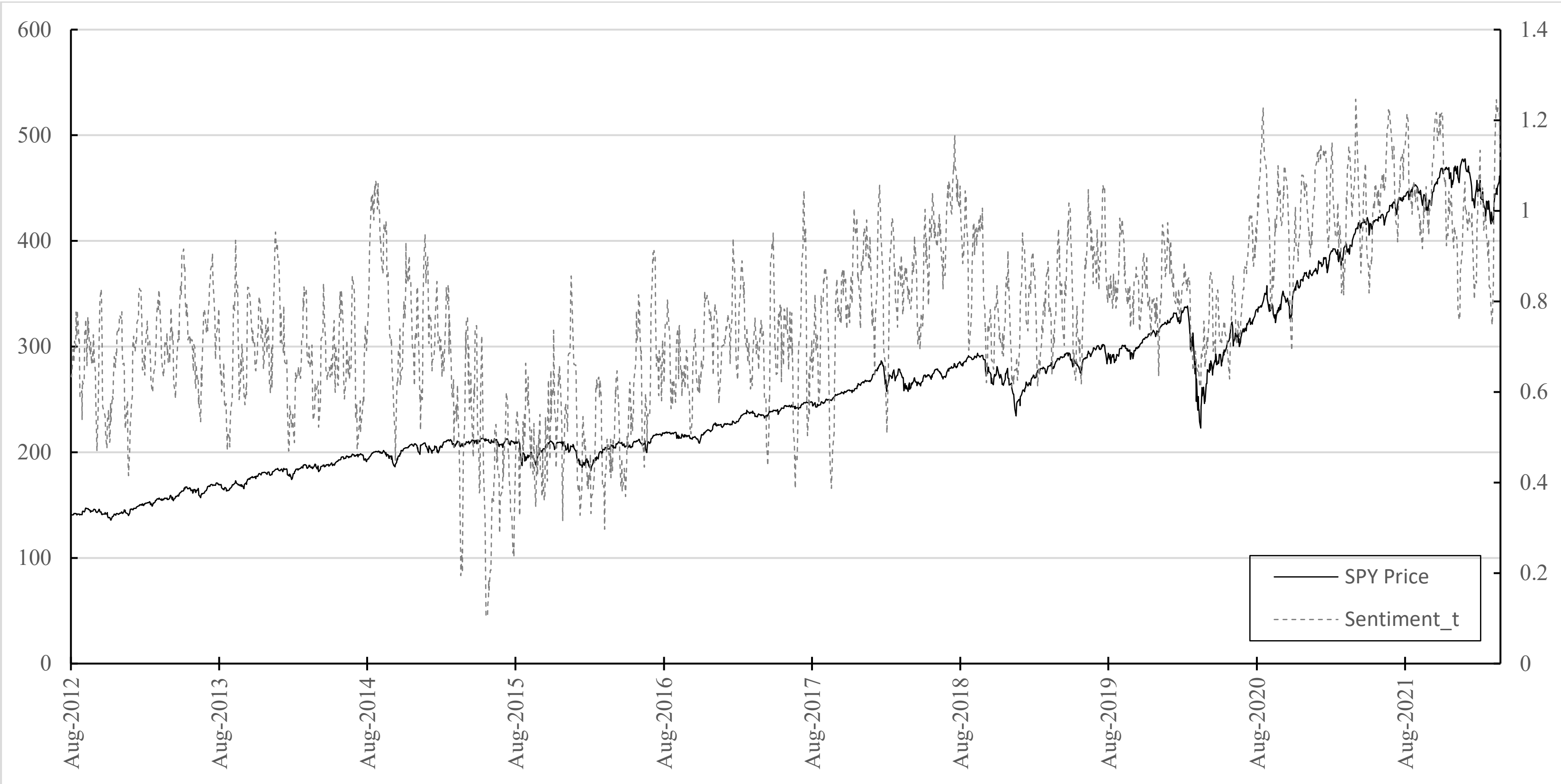

Figure 3. Generalized impulse response from social media sentiment to DH Herding

This figure plots the cumulative impulse response function for one standard deviation shock of the sentiment on the Dollar Based Herding, DH as described in Equation (6). The higher value of Dollar Based Herding means a higher level of herding. The red dotted lines are the 95% upper and lower bands.

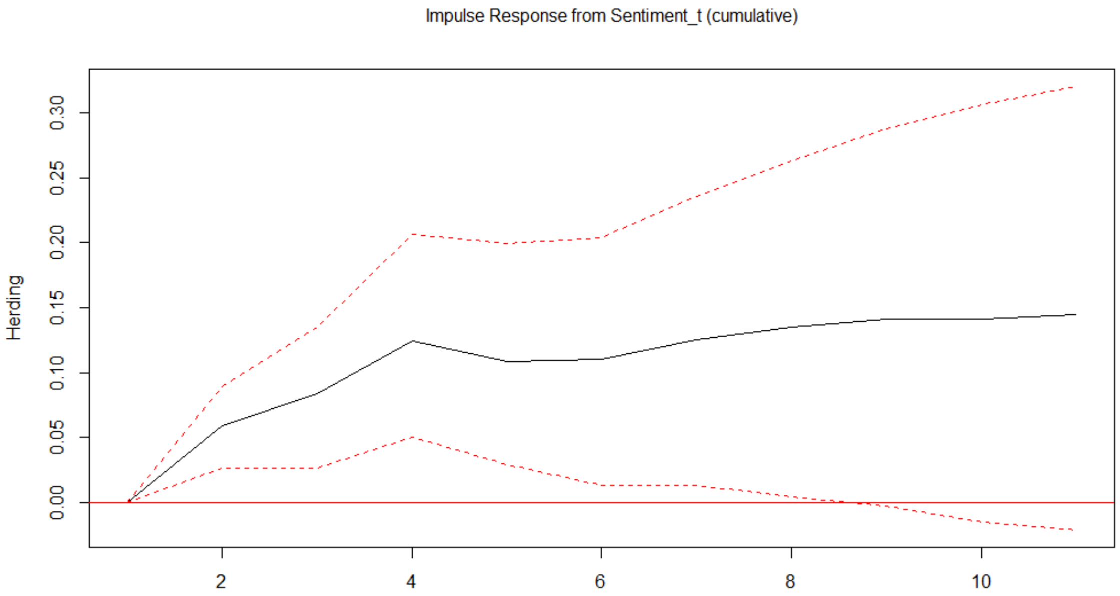

Figure 4. Generalized impulse response from social media sentiment to WR Herding

This figure plots the cumulative impulse response function for one standard deviation shock of the sentiment on the Williams Percent Range, WR, as described in Equation (7). The higher (lower) value of WR means overbought (oversold). The red dotted lines are the 95% upper and lower bands.