Questioning the Wisdom of Crowds: When do social media trends help and hinder retail trading?

March 2024

Abstract

This study examines the impact of retail investors' social network communications on price informativeness. By observing the online forum wallstreetbets on Reddit, which has strong cultural characteristics of creating memes, this study examines under what conditions retail investors' social communication can create informational value in stock trading or mislead other market participants. The results indicate that the more stock tickers were mentioned regardless of the contents of conversations, the higher the price informativeness. Furthermore, when online communications unrelated to firms become more active, stock price informativeness decreases. However, this increased retail investors' general attention to stock trading strengthens the positive impact of firm-specific discussions on price informativeness.

Keywords: Retail investors, Wisdom of crowds, Price informativeness, Social finance, Wallstreetbets

1 Introduction

Technological advances have enabled retail investors to transmit financial information and reprocess it instantly in various distribution formats for other market participants. Retail investors often summarize primary information from firms and governments, distributing it in easily accessible, engaging, and emotionally appealing ways for general public consumption. For instance, on social media platforms such as Reddit, users translate financial information into Internet memes, using comical images and buzzwords to capture the essence of stock price movements, market participants' reactions or macroeconomic trends. These popular "stock memes" have, in turn, generated meme stocks, or stocks that have risen through viral growth in online communications.

The most renowned meme stock occurred in 2021, when floundering stocks in GameStop, a video game company, experienced a dramatic price spike after aggressive promotion online through memes that circulated on the "wallstreetbets" Subreddit. This unprecedented fluctuation in the market resulted in huge losses for hedge funds with short positions, alerting investors to the growing influence of online communication on market stability. The fad of meme stocks and investors' inclination to chase high volatility raise concerns about how the gamification of trading impacts price informativeness, which scholars define as the way that prices reflect the fundamentals of firms. Financial research must therefore further investigate this new trend of viral information transfer to fully understand its implications for market efficiency.

Several studies motivated by the GameStop short squeeze have examined whether online posts and comments shared presumably by retail investors have predictive power for stock returns (for example, Bradley et al., 2021), which would indicate that investors rely on information rather than noise. Yet, these attempts do not clearly explain why online discourse that seems to be closer to entertainment than serious discussion can predict stock returns and how this new trading culture of instant information exchange affects price efficiency. In addition, by tracking the recurrence of stock names or tickers in online forums and correlating this traffic with stock returns, recent studies overlook the possibility that data unrelated to firms in online forums may prevent online users from concentrating on unbiased signals.

Building on these emerging trends in retail trading and related literature, this study investigates the impact of retail investors' social network communications on price informativeness. By observing the online forum, wallstreetbets on Reddit, which has strong cultural characteristics of creating memes, this study examines under what conditions retail investors' casual communication can create informational value in stock trading or mislead other market participants. To reveal this relationship, I web-scraped the textual data from submission posts in wallstreetbets from January 2019 to December 2021, including the timing of the GameStop episode. I then tracked how many times stock names or tickers appeared in the textual data on a firm-month basis, which I call "popularity" in this study, and correlated this traffic with stock price informativeness.

The empirical evidence yields two findings. First, retail investors' attention to stocks helped prices reflect firm-specific information better for the whole sample period, regardless of the quality of the conversations they perceived. This finding was not only robust but also semi-monotonic across various price informativeness and popularity measures; the more stocks were mentioned, the higher the price informativeness. These findings suggest that retail investors in wallstreetbets utilize firm-specific financial data, which is beneficial to market quality in terms of price informativeness. Second, the presence of firm-specific memes is positively correlated to price informativeness. Memes unrelated to a specific stock have a negative impact on price informativeness. However, this increased retail investors' attention amplifies the impact of firm-specific discussions on price informativeness.

This study contributes to two strands of literature that are becoming increasingly important due to the technological advances in the last decade: retail trading and social finance. Specifically, financial innovation introduced by fintech apps and the emergence of social media platforms are notable changes in the trading environment. The advent of fintech applications has led to a decrease in trading commission fees in the overall brokerage industries, which has allowed retail investors to trade with lower costs. The introduction of simple and convenient apps for trading further lowers the barriers to entry for stock investment. Furthermore, various social media platforms such as Twitter, StockTwits, or Reddit allow small and diversified retail investors to actively share their opinion and generate herding behaviors (Barber et al., 2022). These noticeable changes in the recent trading environment reveal the importance of re-investigating changes in retail trading and resulting trading outcomes.

With regard to these changes in retail trading environments, on the one hand, two recent studies shed light on the traits of new retail investors but show mixed results: Barber et al. (2022) and Welch (2021). While both studies focus on the specific type of retail investors called 'Robinhood investors (RH investors)' who are usually inexperienced but active rookie investors in trading, their findings have different conclusions. Barber et al. (2022) demonstrate that RH investors are attention-driven compared to other types of retail investors due to the layout of showing information in the Robinhood app, which results in negative abnormal returns. However, Welch (2022) argues that these investors in aggregate did not underperform compared to standard asset pricing methodologies. He specifically mentions that "Robinhood investors were not collectively 'cannon fodder', exploited by more sophisticated investors elsewhere. Good timing and good stock performance help to explain why RH investors did not attrition out but continued to pour in. (p.1490)" These two mixed results imply that, even if those retail investors are attention-driven rather than fundamental-driven, they may still refer to information as best they can. Regarding this potential mechanism, this study suggest evidence that, retail investors' online communications focusing on entertainment also can help them realize the main news or issues of a specific firm. Also, unlike these two papers that focus on the individual investors' performance, this paper focuses on the market quality by focusing on information contents of communication between retail investors. Eaton et al. (2021) also investigate the impact of RH investors on market quality such as spreads and liquidity, but this paper focuses on price informativeness instead of market quality measures related to liquidity.

On the other hand, in recent discussion of social finance, a key focus is on how investor interaction over social networks affects the outcome of the transaction, which has been accessible to the researcher by observing vast amounts of social network communication data. Additionally, during the COVID-19 pandemic, households have had to enjoy more spare time, which has led to an increase in retail investors' market participation (Ozik, Sadka, and Shen, 2021). These environmental changes make it worthwhile to examine the impact of online communication on the stock market. Hirshleifer (2020) has observed that bias arises when people communicate with each other and transmission of information is incomplete, which affects trading. Although this new view sheds light on why investors' communication is significant, studies about social media communication is at the emerging stage (Bradley et al. 2021; Cookson, Fos, and Niessner, 2021; Hu et al., 2021); researchers have no clue about in what situations this transmission bias is amplified or softened. In this regard, this study contributes to the current literature by showing that the level of aggregate noise in investors' communication amplifies the extent to which the firm-specific information is incorporated into the price.

This study also relates to several studies investigating wallstreetbets and the GameStop episode. Bradley et al. (2021) test the stock return predictability using wallstreetbets submission and comments data and demonstrate that the information shared in wallstreetbets has a predictable power before the GameStop episode. However, they find that the predictable power of the submission post disappears after the GameStop episode. Although the main findings share some insights about retail trading, research motivation and approaches are different with this study. They use only posts classified as 'Due Diligence' that contain investment analysis and recommendation, while this study includes all posts regardless of their classification to disentangle the impact of information-related discussion and entertainment-related discussion. Second, they focus on the return predictability and infer that the Due Diligence posts are informative, while this study directly measures price informativeness and investigates the informational contents of discussion to answer why this conversation increases price informativeness.

The contents of this paper are as follows: Section 2 explains the empirical methodologies suitable for the research questions mentioned above. Section 3 summarizes my main findings for the whole sample period with or without the GameStop example. The structural changes in communication via social networks will be investigated in the next Section 4. Finally, the paper concludes.

2 Methodology

2.1 Sample Selection

Among the common stocks listed on NYSE, NASDAQ, and AMEX, I analyzed only those stocks that have been mentioned more than once by Reddit investors from 2019 to 2021, which is a total of 3,870 companies in the CRSP universe. Some highly mentioned meme stocks - for example, Nokia - are not commons stocks, but the alternative sample selection does not change the main results. The analysis period is from January 2019 to December 2021, and firm-month observations were collected to be analyzed, including the GameStop episode in January 2021. Analysts and earnings information were gathered from I/B/E/S detail history data, and financial information on stocks or firms was downloaded from CRSP and Compustat. The interactions and conversations of retail investors were obtained from the online community called wallstreetbets, one of the subreddits specialized for exchanging investment opinions in easily accessible and emotionally attractive ways. Official tickers were used to link this textual data to financial data from CRSP on a firm basis, which will be further explained. The sample size will be varied according to the availability of financial variables during the window and various popularity measures.

2.2 Variable Construction

The two most important variables in this study are *popularity* and *price informativeness*. The variable, *popularity*, is the number of times wallstreetbets users have mentioned a stock ticker in text data, which captures the amount of interest the stock has received from retail investors. The variable, *price informativeness*, was mainly constructed by transforming the R-squared of the market model. The proxy for market returns was daily CRSP value-weighted returns, and returns for industry portfolios and Fama and French three factors were also considered to disentangle industry, size and value effect. Other financial variables associated with these two variables were included to control potential effects.

2.2.1 Popularity

The monthly variable, popularity, is defined by the number of times a specific stock ticker is mentioned in r/wallstreetbets during a month. It can also be interpreted that this measure is the intensity of investors' communication on stocks. Three steps are required to construct popularity.

Step 1. Collect the textual data from wallstreetbets

To extract the information needed for this study, JSON files for submission posts and comments were downloaded from a third-party API called Pushshift from January 2019 to December 2021 and processed by the JQ processor: the unique ID assigned to each post or the ID of the post where the comment was posted, author's username, URL, Unix time when the post/comment was submitted, title/selftext/body, link flair class/text, and score. In sum, 1,775,888 posts were extracted for 2019.01-2021.12. Table 1 shows the number of extracted posts for each month. According to the table, during the GameStop episode in January, the number of submitted posts was extremely high, compared to before and after the episode. The number of entire posts has increased about 15 times between December 2020 and January 2021.

Month	Year 2019	Year 2020	Year 2021
01	6,230	11,185	536,691
02	4,408	28,310	398,113
03	4,660	48,935	125,046
04	6,002	26,851	49,730
05	5,619	20,823	44,765
06	5,089	24,203	94,987
07	5,576	21,993	25,759
09	8,883	23,535	25,620
10	5,364	16,833	25,781
11	5,052	28,228	28,368
12	7,604	32,623	21,124

Table 1: The number of submission posts for 2019-2021

The unique ID of the post allows us to distinguish between posts with different content under the same title. The URL provides a link to access extracted posts directly from the website, and remains the same when deleted. Unix time contains information in seconds regarding when the post/comment was submitted. It converted to a readable time display based on Estern Time (EST/EDT). The score is a number that subtracts downvote from upvote, indicating the user's affinity for the post/comment. Titles, selftexts, and bodies contain textual data, varying in length. Some submission posts have been removed by moderators due to Reddit's policy (e.g., classified as spam) or later self-deleted, in which case their own text has been marked marked "[removed]" or "[deleted]." Even in this case, the initial version of the title and its URL remain the same. The number of deleted posts which clearly contain "[removed]" or "[deleted]" in selftext is 1,162,489. There are 478,894 cases in which there is no text in the textbody (i.e., empty). Put together, about 95.9% (=1,641,383/1,711,401) of submission posts will be excluded from the analysis if we only consider the complete cases that have both the title and text body. For this reason, to contain as much data as possible, the title of the post was analyzed mainly.

Moreover, there were cases that the posts had exactly the same title. One common case is that the title consists of general and easily repeatable words such as 'daily discussion'. In this case, to capture the impact of this type of posts, the similar analysis was performed by considering the contents of posts. Another common case is that the same author uploaded the duplicated post with the same title after the previous post was deleted from moderators. The unique ID was also given to this duplicated post although the title, author, contents were the same. It is relatively complicated to determine whether this information is new or old on the user's side because the posts that were uploaded again can be exposed to other distinct users. The cases that the comments under the newly uploaded duplicated post were different from the previous one support this possibility. To determine the uniqueness of the post depends on where the analysis focuses. This study applies the same regression methodologies to both cases in analyses: including all titles and excluding the duplicated titles written by the same author. In sum, the duplicated posts account for about 4% of the total sample. The number of duplicated posts proportionally increased when the number of entire submissions increased, especially after the GameStop episode. This phenomenon may indicate the existence of the echo-chamber effect among individual investors through the social network as in the theoretical model by Pedersen (2022).

Step 2. Handle and tokenize the textual data

According to Reddit users' writing patterns, the collected titles were handled in a direction that can minimize the noise in text, so that the number of times the tickers mentioned can be clearly identified. Then, all titles were tokenized (i.e., the sentence was separated as the collection of words) and compared with stock tickers.

First, Reddit users tend to put special characters like '$' in front of tickers. Thus, if the text includes '$' like '$GME', the algorithm that matches the tokens with the stock tickers count this case. Other than '$', all other special characters were removed or replaced with spaces for the following reasons. Usually, deleting special characters should be carefully considered because removing special characters can distort the meaning of words such as 'I'm' and 'GME & AMC'. This study is relatively free from this concern since the stock ticker consists solely of alphabets but using an apostrophe to represent possessive like 'GME's' or 'I'm' may cause the problem. If the apostrophe is just deleted, then the word changes to 'GMEs', thus the algorithm cannot count this case. On the other hand, if the apostrophe is replaced with spaces, one-word changes to two words as 'GME' and 's'. This replacement causes the number of times the ticker 'S' (SentinelOne Inc.) mentioned to be surprisingly high. In this analysis, the special characters were replaced with spaces, which implies that a 1-unit digit ticker should be carefully treated: in case of a 1-unit digit ticker, I used '$' to certainly detect the stock ticker only.

Another situation requiring a text-handling process is when the title includes emojis. These emojis lead to a similar situation as special characters do, so they were also replaced with spaces. All strings in the title and textbody became lowercase to avoid any confusion that capital letters might bring; 'GME' and 'gme' are identically treated. Finally, the lowercase title and textbody without special characters and emojis were combined as the refined title. Each refined title was tokenized for counting. To be more specific, every word in the title was separated as one respective object. For instance, if we assume that the original title is '$GME to the moonâ€™.', the tokenized refined title becomes 'gme', 'to', 'the', and 'moon'. The number of spaces between words does not exert any influence in the tokenization process; ' GME' is regarded as 'gme'. Clearly, we can now conclude that GME was mentioned once in this example. As this example shows, the tokenized words have advantages in terms of counting but are disadvantageous of interpreting the context associated with the stock ticker (e.g., why the stock ticker was mentioned by users).

Step 3. Build the variable, *popularity*, on the monthly basis

After collecting the refined titles, the number of times the stock was mentioned was counted to build the variable, *popularity*. Even after refining the texts, the noise still exists due to the 1-unit alphabet tickers and the stop words in the counting process. Usually, stop words in the text processing indicate the words such as 'by' or 'on' which has no specific meaning in the context. In the wallstreetbets' textual data, these stop words are highly likely to be used as preposition rather than tickers. To reduce the noise caused by these stop words, *popularity* excludes both the 1-unit alphabet tickers and stop words. Thus, the words such as 'for' or 'all' were excluded from the analysis, except for the cases that it included '$'

in front of the stop words. *popularity* still has some words that can have their own meaning as a word such as 'CAN', 'NOW', and 'GO'. If we collectively drop all tickers that have their own function as words, this exclusion may lead to lose information contained in the cases that the words actually indicate the ticker. Thus, in this case, this paper only considers the cases where '$' was tagged in front of the words.

Finally, I added 1 to the aggregated number after counting in order to include cases where the company ticker was mentioned during the three-year analysis period but not in that month. Also, due to its skewed distribution, the natural logarithm was taken for each popularity measure. Put differently, *popularity* is defined as log(1+ # times a ticker has been mentioned).

2.2.2 Price informativeness

According to Roll (1988), R-squared obtained from the market model has been transformed and used to measure *price informativeness*, which is also called price asynchronicity (for example, Piotroski and Roulstone, 2004; Morck et al., 2013). The conventional market model is

$$R_{i,t} = \alpha_0 + \alpha_1 R_{M,t} + \epsilon_{i,t} \tag{1}$$

where $R_{i,t}$ is the individual stock return and $R_{M,t}$ is the market return. After obtaining R^2 from the market model, *price informativeness* is constructed by:

$$\mathit{prcinfo} = \log\left(\frac{1-R^2}{R^2}\right) \tag{2}$$

In the market model, R-squared represents how much the variation of the market return explains that of individual stock returns. The low R-squared means that the market return variation less explains the stock return variation, which implies that the firm-specific information is more incorporated into the individual stock return. Thus, relatively low R-squared suggests a high level of price informativeness.

There has been debate in the literature on this transformation of R^2 represents price informativeness (Piotroski and Roulstone, 2004). The process of constructing this measure naturally assumes that there are two types of information, market-wide and firm-specific, which overlook the possibility that the remaining variation may be caused by firm-specific noise. Despite this limitation, this study mainly uses Equation (2) as a proxy for *price informativeness*. This is mainly because of the distinctive features of wallstreetbets data; we can collect interaction and conversation data, which may contain information about stocks, *in seconds*. Earning announcements are usually known as firm-specific information, but these are announced at least every quarter. In contrast, Equation (2) is not limited to a certain time frequency. Especially, using quarterly data may miss the impact of unprecedented events such as the GameStop episode. To capture firm-specific information as much as possible, returns on industry portfolios and Fama and French three-factor portfolios were also included in the market model.

2.2.3 Other control variables

All control variables are obtained from CRSP, I/B/E/S, and Compustat. Except for the bi-weekly reported short interest, any other variables are updated on the daily basis. To perform a monthly regression, all variables were converted to the monthly unit by keeping the end-of-month data. *volume* is the trading volume which not only reflects the case where investor's attention leads to a transaction but also represents the stock liquidity. *mktcap* indicates the current market capitalization (i.e., price multiplied by current shares outstanding) as a proxy of firm size. *volatility* is the 30-day volatility obtained. For the newly listed firms that do not have enough observations for the first 29 days, volatility is a missing value. *siratio* is the short interest (i.e., the total number of shorted shares that have not yet been bought back) divided by the total number of shares outstanding, which is pointed out as the main factor attracting Reddit investors.

2.3 Descriptive Statistics

Table 2 summarizes the statistics of key variables for 2019.01-2021.12. *prcinfo* where R-squared is obtained from the market model. *popularity* is defined as the number of times the stock tickers were mentioned in wallstreetbets for the entire posts. *mktcap* (market capitalization), *siratio* (short interest ratio), *volume* (trading volume), and *vol30* (30-day volatility) are based on end-of-month value and taken by natural logarithm. Note that the average of *popularity* is close to its minimum suggesting that Reddit users' interest was unequally distributed and focused on some popular stocks. Specifically, the *popularity* minimum is 1 (log1=0), which implies that certain tickers are mentioned once a month. In contrast, the maximum popularity is 11.3936, which is the popularity of GameStop (GME) in January (log(88753+1)=11.3936).

Table 3 shows the further investigation of the number of times tickers were mentioned. 67% (=76,896/114,068) of the total sample has never been mentioned in a certain month. In contrast, only 1.4% (=1,630/114,068) of the total sample has been mentioned more than 100 times in a month, and even in this group, this measure is extremely widely distributed from

Variable	Obs	Mean	Std. dev.	Min	Max
# times the ticker mentioned	145,938	11.8551	357.4600	0.0000	88753.0000
popularity	145,938	0.5526	1.0816	0.0000	11.3936
prcinfo	145,938	2.2442	2.1361	-2.6311	24.3589
size	145,938	13.6575	2.2724	4.4144	21.7686
volume	145,938	12.9161	1.8275	-1.1896	20.3575
volatility	145,938	-3.5397	0.7075	-7.6547	0.8174
siratio	145,876	-4.0991	1.8272	-18.4161	0.0000
analyst coverage	145,938	0.9076	0.9295	0.0000	3.8712

Table 2: Descriptive Statistics

100 to 88,753. To mitigate the potential problem that may be caused by this distribution, even after taking a log, various specifications will be used in the next section.

# Mention	Obs	Mean	Std.	Min	Max
=0	76,896	0	0	0	0
=1	13,367	1	0	1	1
>1 & <10	15,253	3.8899	2.0592	2	9
>10 & <100	6,922	29.6581	21.0596	10	99
>100	1,630	638.0172	3,079.3330	100	88,753
Total	114,068	11.5542	375.7442	0	88,753

Table 3: The number of times the ticker has been mentioned by group

2.4 Empirical Strategies

2.4.1 Main specification

In this section, the relationship between *popularity* and *price informativeness* will be explored. If the stock is popular due to its financial information that retail investors evaluate as potential price changes, their active communications (i.e., high popularity through the social network) allow that specific information to be incorporated into price well. To empirically test how stock popularity affect price informativeness, the below regression will be performed. Based on Chan and Hameed (2006) and by considering that short interest ratio may be the characteristics that Reddit investors prefer, the empirical model is the following:

$$\begin{aligned} \mathit{PrcInfo}_{i,t} &= \alpha + \beta_1 \mathit{LOG(1{+}popularity}_{i,t}) + \beta_2 \mathit{LOG(size}_{i,t}) + \beta_3 \mathit{LOG(volume}_{i,t}) \\ &+ \beta_4 \mathit{LOG(coverage}_{i,t}) + \beta_5 \mathit{LOG(siratio}_{i,t}) + \beta_6 \mathit{LOG(volatility}_{i,t}) + \theta_i + \epsilon_{i,t} \end{aligned} \tag{3}$$

The most noteworthy coefficient is β_1, which shows how popularity relates to price informativeness. The results were obtained by performing the panel regression with fixed effects and firm cluster standard errors to control the firms' unobservable characteristics that may affect Reddit investors' preferences.

2.4.2 Alternative specifications

As Table 3 suggests, even if the log was taken to adjust its distribution, there may be a potential problem caused by the fact that the variable, *popularity* has a value concentrated below 10. To see if the main results change due to a small number of famous meme stocks such as GME or AMC, the following additional specification are also considered.

$$\begin{aligned} \mathit{PrcInfo}_{i,t} &= \alpha + \beta_1 \mathbb{I}(\text{Covered}) + \beta_2 \mathit{LOG}(\mathit{size}_{i,t}) + \beta_3 \mathit{LOG}(\mathit{volume}_{i,t}) \\ &+ \beta_4 \mathit{LOG}(\mathit{coverage}_{i,t}) + \beta_5 \mathit{LOG}(\mathit{siratio}_{i,t}) + \beta_6 \mathit{LOG}(\mathit{volatility}_{i,t}) + \theta_i + \epsilon_{i,t} \end{aligned} \tag{4}$$

where $\mathbb{I}$(Covered) is an indicator variable, which gives 0 to the stock tickers never mentioned in a month, otherwise. The coefficient β_1 captures the effect when a stock that has never been mentioned is mentioned. Additionally, I divided the sample that had been mentioned more than twice into five quintiles to see what additional effects of being covered by retail investors:

$$\begin{aligned} \mathit{PrcInfo}_{i,t} &= \alpha + \beta_1 \mathbb{I}(\text{One-time}) + \beta_2 \mathbb{I}(\text{P20}) + \beta_3 \mathbb{I}(\text{P40}) + \beta_4 \mathbb{I}(\text{P60}) \\ &+ \beta_5 \mathbb{I}(\text{P80}) + \beta_6 \mathbb{I}(\text{P100}) + \delta\,(\mathit{Controls})_{i,t} + \theta_i + \epsilon_{i,t} \end{aligned} \tag{5}$$

where $\mathbb{I}$(One-time) is an indicator variable that gives 1 to the stock tickers mentioned only once in a month, otherwise. $\mathbb{I}$(P20) is the one that gives 1 to the stock if the ticker belongs to the first quintile, and $\mathbb{I}$(P100) is the one that gives 1 to the stock if the ticker belongs to the last quintile.

3 Main Results

This section shows the empirical results described above. Table 4 describes that how *price informativeness* changes according to the level of popularity. Column (1) shows that, even after controlling the impacts of other control variables, the prices of the stocks that have been mentioned more that once are more informative than those that never mentioned in certain month. This finding is consistent across various price informativeness measures; *price informativeness* in Column (2), (3), and (4) were obtained by including market returns, market and industry returns, and FF3 portfolio returns, respectively.

VARIABLES	(1) *prcinfo1*	(2) *prcinfo1*	(3) *prcinfo2*	(4) *prcinfo3*
$\mathbb{I}$(Covered)	0.0688***			
	[4.40]			
popularity		0.0701***	0.0651***	0.0469***
		[5.81]	[7.26]	[7.35]
size	-0.3350***	-0.3434***	-0.2745***	-0.2191***
	[-18.52]	[-18.89]	[-18.39]	[-22.12]
volume	0.1214***	0.1160***	0.1585***	0.1345***
	[9.40]	[9.02]	[14.88]	[19.41]
volatility	0.2078***	0.1990***	-0.1299***	-0.1093***
	[8.16]	[7.78]	[-7.50]	[-9.22]
siratio	-0.1692***	-0.1692***	-0.0932***	-0.1185***
	[-16.10]	[-16.13]	[-11.20]	[-21.19]
analyst coverage	0.0743***	0.0730***	0.1929***	0.1797***
	[9.84]	[9.67]	[30.26]	[36.86]
Constant	4.3750***	4.5195***	1.4450***	0.8558***
	[14.26]	[14.72]	[5.60]	[5.20]
Firm FE	Yes	Yes	Yes	Yes
Year-Month dummy	Yes	Yes	Yes	Yes
Observations	114,031	114,031	83,539	113,782
R-squared	0.148	0.148	0.158	0.171
Number of firms	3,870	3,870	2,639	3,865

Table 4: The impact of popularity on price informativeness

VARIABLES	(1) *prcinfo1*	(2) *prcinfo1*	(3) *prcinfo2*	(4) *prcinfo3*
$\mathbb{I}$(Covered)	0.0689***			
	[4.41]			
popularity		0.0701***	0.0784***	0.0645***
		[5.77]	[6.44]	[7.15]
Control	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Year-Month dummy	Yes	Yes	Yes	Yes
Observations	113,995	113,995	113,750	83,503
R-squared	0.148	0.148	0.150	0.158
Number of firms	3,869	3,869	3,864	2,638

Table 5: The impact of popularity on price informativeness (without GME)

Table 5 shows how price informativeness changes according to the level of popularity when the outlier, GameStop, is excluded from the sample. The results based on the whole sample periods are consistent with previous findings. Table 6 shows the estimation results from Equation (5), which mitigats the concerns arising from the widely distributed popularity. Consistent with the previously findings, every coefficient of dummy variables is positive and significant. This finding was not only robust but also semi-monotonic across various price informativeness; The more stocks were mentioned, the higher the price informativeness.

VARIABLES	(1) *prcinfo1*	(2) *prcinfo2*	(3) *prcinfo3*
$\mathbb{I}$(One-time)	0.0417**	0.0322*	0.0629***
	[2.38]	[1.70]	[4.73]
$\mathbb{I}$(P20)	0.0939***	0.0414	0.0773***
	[3.51]	[1.49]	[3.88]
$\mathbb{I}$(P40)	0.0924***	0.1322***	0.1265***
	[3.31]	[4.09]	[5.71]
$\mathbb{I}$(P60)	0.1470***	0.1535***	0.1228***
	[3.95]	[4.12]	[4.78]
$\mathbb{I}$(P80)	0.2010***	0.2055***	0.1612***
	[4.64]	[4.70]	[5.08]
$\mathbb{I}$(P100)	0.2828***	0.3366***	0.2395***
	[4.82]	[5.77]	[5.35]
size	-0.3426***	-0.3915***	-0.2725***
	[-18.83]	[-22.45]	[-18.29]
volume	0.1161***	0.1453***	0.1585***
	[9.01]	[11.33]	[14.89]
volatility	0.1994***	0.0611***	-0.1288***
	[7.81]	[2.72]	[-7.44]
siratio	-0.1695***	-0.1587***	-0.0936***
	[-16.14]	[-16.14]	[-11.24]
analyst coverage	0.0729***	0.1971***	0.1923***
	[9.65]	[23.10]	[30.08]
Firm FE	Yes	Yes	Yes
Year-Month dummy	Yes	Yes	Yes
Observations	114,031	113,786	83,539
R-squared	0.148	0.150	0.158
Number of permno	3,870	3,865	2,639

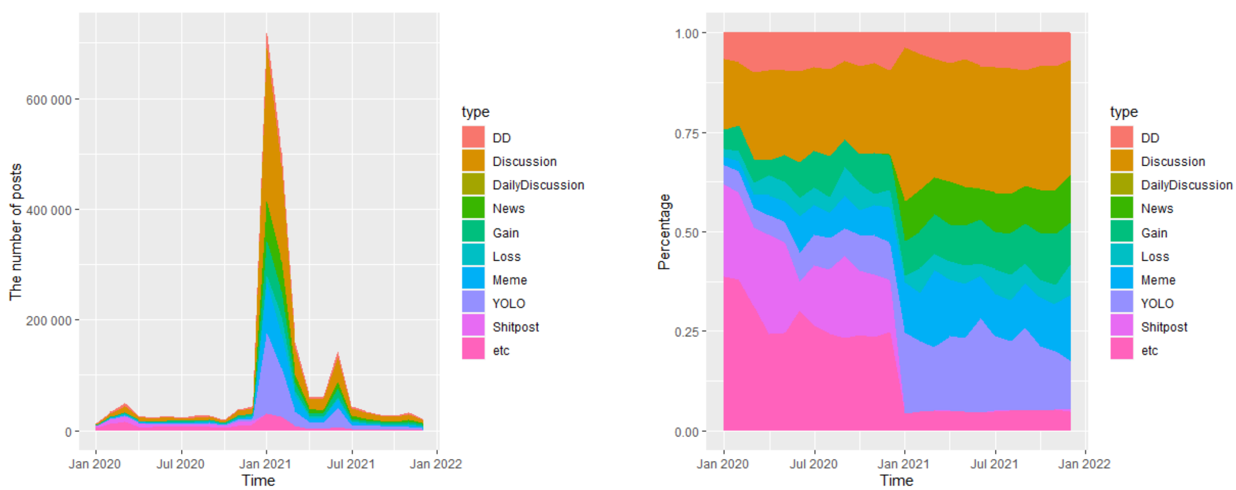
Table 6: Quantile regression: the impact of popularity on price informativeness

4 Potential Channels: Information vs. Noise

Given wallstreetbets' strong cultural characteristics of creating memes, how and why does stock price informativeness increase when users mention the specific firms more frequently? To answer these questions, all posts are classified based on the subject of discourse and the contents of discourse. More specifically, posts are categorized as firm-specific or non-firm-specific (*the subject of discourse*) and as information or entertainment (*the contents of discourse*). Thus, posts can be (1) firm-specific information, (2) non-firm specific information, (3) firm-specific entertainment contents, or (4) non-firm-specific entertainment contents. In this section, I examine how the posts in each category affect price informativeness to distinguish the impact of informational contents from that of entertainment contents.

4.1 The subject and contents of discourse

I identified two categories of submission posts using the 'link flair text' tags assigned by the authors: those likely to be associated with information (e.g., 'DD', 'Discussion', 'Daily Discussion' and 'News') and those likely to be related to entertainment (e.g., 'Meme', 'YOLO', and 'Shitpost').



(a) The number of postings by group

(b) Proportions of postings by group

Figure 1: Comparison of posting statistics

Figure 1 summarizes the number of posts that I counted in each of the nine categories, sorted in descending order, and the proportion of posts for each category. Figure 1.(a) illustrates the significant increase in the number of postings during the GameStop episode, while Figure 1.(b) shows a substantial rise in posts classified as entertainment after the GameStop event. The proportion of 'DD' and 'Discussion,' which represents informative content, remains relatively unchanged. This graphical change in proportion may imply that even within the same communication platform, users' habits of sharing and consuming data have changed in ways that can affect price informativeness. After the GameStop event, users did not return to the previous mode of communication: entertainment-related posts with short text and comical images continued to attract more attention. However, the proportion of posts regarding their performance shows that users remain more willing to talk about their 'gains' than their 'losses.'

I have also divided the information and entertainment posts into two further categories: firm-specific and non-firm-specific posts. I identified firm-specific posts as those including a specific firm ticker and possibly containing either information or entertainment content. In contrast, I defined non-firm-specific posts as covering a broader range, including macroeconomic information, questions about general stock investment, and images with no context. These posts can also be classified as either informational or entertainment but do not contain the ticker of specific firms.

Posts that are categorized as firm-specific information are expected to significantly and positively affect price informativeness. In contrast, non-firm-specific entertainment content is expected to prevent users from concentrating on valuable information, resulting in a decrease in price informativeness. However, these non-firm-specific entertainment posts induce users to stay longer on the platform and be exposed to information included in the viral trends. A firm-specific meme can be a signal to users that something is happening about the firm. If this signal prompts users to search further for firm-specific information, then the price informativeness increases. Otherwise, users may also be distracted by entertainment content. The impact of firm-specific entertainment content on price informativeness should be investigated by looking at the data.

4.2 Model specification

To investigate the possibility that price informativeness increases when there is more non-firm-specific noise, I use the following regression.

$$\begin{aligned} (\mathit{PrcInfo})_{i,t} &= \beta_0 + \beta_1 (\mathit{Info\_popularity})_{i,t} + \beta_2 (\mathit{memed})_{i,t} \\ &+ \beta_3 (\mathit{Info\_popularity} \times \mathit{memed})_{i,t} + \delta (\mathit{Controls})_{i,t} + \theta_i + \epsilon_{i,t} \end{aligned} \tag{6}$$

Info_popularity is the number of times a stock mentioned in posts categorized as discussion, such as 'DD', 'Discussion', and 'Daily Discussion', etc. *Memed* takes a value of 1 if the monthly number of meme posts for a stock is greater than the average monthly number of meme posts (categorized as 'Meme', 'YOLO', and 'Shitpost') across all stocks, and 0 otherwise. It is expected for the coefficient of *Info_popularity* to be positively significant because this variable measures the number of posts classified as information.

VARIABLES	*PrcInfo*
Info_popularity	0.0996***
	[6.27]
memed	0.0847*
	[1.71]
Info_popularity × *memed*	-0.0091
	[-0.37]
Controls	Yes
Firm FE	Yes
Year-Month dummy	Yes
Observations	96,672
R-squared	0.148

Table 7: The different role of information and meme I

The posts can also be categorized into two types: firm-specific and non-firm-specific posts. Firm-specific posts including a specific firm ticker may contain either informational or entertainment content, which can be inferred from the 'link flair text'. Non-firm-specific posts include a wider scope, such as macroeconomic information, users' investing performance, or questions about general stock investment. The following regression equation is to study if price informativeness increases when there are more non-firm-specific noise.

$$\begin{aligned}(\mathit{PrcInfo})_{i,t} &= \beta_0 + \beta_1(\mathit{Info\_popularity})_{i,t} + \beta_2(\mathit{noise\_level})_t \\ &+\beta_3(\mathit{Info\_popularity} \times \mathit{noise\_level})_{i,t} + \delta(\mathit{Controls})_{i,t} + \theta_i + \epsilon_{i,t}\end{aligned} \quad (7)$$

where *noise_level* captures monthly variations in the number of non-firm-specific meme posts and takes a value of 1 if the monthly number of these memes exceeds the median.

4.3 Results

Table 7 shows that on average, firms with a higher frequency of memes have higher price informativeness, although the relationship is only marginally significant. Table 8 shows that, when the aggregate level of noise is greater, price informativeness decreases. Increased retail investor attention amplifies the impact of firm-specific discussion on price informativeness.

VARIABLES	*PrcInfo*
Info_popularity	0.1111***
	[6.19]
noise_level	-0.2481***
	[-14.32]
Info_popularity × *noise_level*	0.0355**
	[2.51]
Controls	Yes
Firm FE	Yes
Year-Month dummy	Yes
Observations	96,672
R-squared	0.023

Table 8: The different role of information and meme II

5 Conclusion

The trading activity of individual investors in the stock market has long been investigated. The previous studies have explored various aspects of retail traders' market participation, such as their performance (Barber and Odean, 2000), their roles as the noise traders (Barber and Odean, 2008; Mendel and Shleifer, 2012), the low market participation of retail investors (Campbell, 2006), and the heterogeneity in investing skills among retail investors (Coval, Hirshleifer, and Shumway, 2005; Korniotis and Kumar, 2013). Despite these academic efforts to understand individual investors' market participation, the noticeable changes in the recent trading environment require re-investigating the individual investors' role.

This study decomposes individual investors' online stock discussions into two dimensions and then examines how each information content affects price information. The number of times a stock is mentioned by Reddit investors has a positive impact on its price informativeness. The presence of firm-specific memes is also positively associated with price informativeness. However, the aggregate noise, non-firm-specific memes has a negative impact on price informativeness.

References

[1] Barber, B. M., Huang, X., Odean, T., & Schwarz, C. (2022). Attention-induced trading and returns: Evidence from Robinhood users. *The Journal of Finance*, 77(6), 3141-3190.

[2] Barber, B. M., & Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *The Journal of Finance*, 55(2), 773-806.

[3] 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.

[4] Bradley, D., Hanousek Jr, J., Jame, R., & Xiao, Z. (2021). Place Your Bets? The Market Consequences of Investment Advice on Reddit's Wallstreetbets. *Working Paper*.

[5] Campbell, J. Y. (2006). Household finance. *The Journal of Finance*, 61(4), 1553-1604.

[6] Chan, K., & Hameed, A. (2006). Stock price synchronicity and analyst coverage in emerging markets. *Journal of Financial Economics*, 80(1), 115-147.

[7] Cookson, J. A., Fos, V., & Niessner, M. (2021). Does Disagreement Facilitate Informed Trading? Evidence from Activist Investors. Evidence from Activist Investors. *Working Paper*.

[8] Coval, J. D., Hirshleifer, D. A., & Shumway, T. (2005). Can individual investors beat the market?. *Working Paper*.

[9] Eaton, G. W., Green, T. C., Roseman, B., & Wu, Y. (2021). Zero-commission individual investors, high frequency traders, and stock market quality. High Frequency Traders, and Stock Market Quality. *Working Paper*.

[10] Hirshleifer, D. (2020). Presidential address: Social transmission bias in economics and finance. *The Journal of Finance*, 75(4), 1779-1831.

[11] Hu, D., Jones, C. M., Zhang, V., & Zhang, X. (2021). The rise of reddit: How social media affects retail investors and short-sellers' roles in price discovery. *Working Paper*.

[12] Korniotis, G. M., & Kumar, A. (2013). Do portfolio distortions reflect superior information or psychological biases?. *Journal of Financial and Quantitative Analysis*, 48(1), 1-45.

[13] Mendel, B., & Shleifer, A. (2012). Chasing noise. *Journal of Financial Economics*, 104(2), 303-320.

[14] Morck, R., Yeung, B., & Yu, W. (2013). R-squared and the economy (No. w19017). National Bureau of Economic Research.

[15] Ozik, G., Sadka, R., & Shen, S. (2021). Flattening the illiquidity curve: Retail trading during the COVID-19 lockdown. *Journal of Financial and Quantitative Analysis*, 56(7), 2356-2388.

[16] Pedersen, L. H. (2022). Game on: Social networks and markets. *Journal of Financial Economics*, 146(3), 1097-1119.

[17] Piotroski, J. D., & Roulstone, D. T. (2004). The influence of analysts, institutional investors, and insiders on the incorporation of market, industry, and firm-specific information into stock prices. *The accounting review*, 79(4), 1119-1151.

[18] Roll, R. (1988). Presidential address: R2. *The Journal of Finance*, 43(2), 51-566.

[19] Welch, I. (2022). The wisdom of the Robinhood crowd. *The Journal of Finance*, 77(3), 1489-1527.