Essays in Asset Pricing: Social Media, Retail Trading and Uncertainty

Yue Ma

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:
Eric Zivot, Chair
Quan Wen
Mark Westerfield

Program Authorized to Offer Degree:
Economics

University of Washington

Abstract

Essays in Asset Pricing: Social Media, Retail Trading and Uncertainty

Yue Ma

Chair of the Supervisory Committee:
Chair Eric Zivot
Department of Economics

Have you ever wondered whether it is worth your time to check social media for stock trading ideas? I have, and this dissertation is dedicated to understanding retail traders, social media, and uncertainty. It consists of three chapters that address questions around (1) the information content of social media (Reddit *r/WallStreetBets*) and the role of Reddit influencers, (2) the role of Robinhood retail traders in the stock market and provide an explanation for the inconsistencies in the literature, and (3) the relationship between two types of retail traders (Robinhood traders vs. TAQ traders), their interactions, and how social media affects retail trading.

Chapter 1 aims to provide an understanding of the social media Reddit, particularly the subreddit *r/WallStreetBets* (WSB). In this chapter, I describe the data scrapping process and construction of various Reddit WSB activity measurements (daily discussion attention, all attention, sentiment based on Loughran and McDonald (2011), and sentiment based on Hutto and Gilbert (2014)). I further analyze the information content of Reddit WSB that entails (1) stock return predictability of WSB activities, and (2) whether Reddit WSB users respond to the firm's fundamental information, proxied by firms' key developments. At the end of this chapter, I explore how Reddit WSB influencers affect the subreddit under the theoretical guidance of Pedersen (2021). My findings suggest that Reddit WSB can predict stock returns over a 10-day period and respond to most of the firms' key development events.

Contrary to the common belief, the Reddit WSB influencers' activities do not predict the attention/sentiment of the entire subreddit. This chapter adds to the literature on social media, social networks, and retail trading.

The second chapter of my dissertation provides theoretical and empirical evidence on the impact of Robinhood (RH) traders on stock uncertainty[1]. We find that RH traders increase idiosyncratic risk but decrease total volatility. We are able to bridge contradictory findings in the literature on the effect that RH traders have on total and idiosyncratic volatility. We support our findings by showing that, in general, RH traders react to firm-specific news, which increases idiosyncratic risk, a risk that can easily be diversified away. Finally, we show that sentiment derived from the Reddit social media platform explains a significant portion of the investment decisions of RH traders.

Lastly, chapter 3 aims to provide more stylized facts about the similarities and differences between RH and TAQ traders. TAQ traders are the general retail traders that can be identified in the NYSE Trade and Quote (TAQ) database, based on the algorithm proposed in Boehmer et al. (2021). My findings confirm that RH and TAQ traders respond to information on social media differently, as many market surveys suggested. I am also interested in the interaction between different types of retail trading, as Pedersen (2021) suggested, rational investors can follow fanatic trading under some circumstances to make profits. I find a weak relationship between TAQ and RH net buying. When zooming in on a group of popular stocks on either RH or Reddit, RH net buying is negatively related to stock return, and positively related to firms' idiosyncratic volatility, while the TAQ marketable order imbalance is on the opposite, establishing a positive relationship with stock return and negative relationship with idiosyncratic volatility. These findings indicate that RH traders exhibit more noise trading characteristics, compared to TAQ traders.

---

[1]Joint work with Anthony Sanford, HEC Montréal, Assistant Professor

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First of all, my deepest gratitude goes to my supervisory committee chair, Eric Zivot, thank you for the bi-weekly meetings, your encouragement, and guidance. I have learned a lot from you! I would also like to thank my committee members, Quan Wen, Mark Westerfield, and Jennifer Koski, thank you for your patience and encouragement while I navigate through mountains of papers and thoughts. I am very grateful that I have you on my committee. Last but definitely not least, thank you, Anthony Sanford, for being the "$p < 0.01$" coauthor, I enjoy discussing research ideas with you, thank you for being an inspiration and for all the help you've given to me.

My husband, Yida, believed in me throughout my doctoral process. Thank you for introducing the idea of getting a Ph.D. to me, and trusting my decision to pursue it, even if that means you have to see me 24/7, in the graduate student's office and at home. Thank you, I had lots of fun! Also, thank you for the long lakeside walks and all the debates about whether macroeconomics (his field) or microeconomics is more interesting and useful. For the record, I won. My parents, since day one of my Ph.D., give me their full support. I am forever grateful for the love and trust they have for me, and for helping me to see how beautiful the world is. You are my motivation to be a better person. Thank you, thank you, thank you!

Lia and Zhou, thank you for answering my random late-night econometrics questions, and for all the spontaneous trips we've made this year. Zhe, thanks for being a wonderful friend and wish you success every step of the way. Herbert, Darwin, Hae Yun, Jackson, and Sajid, we started together as a cohort and that will always have a special place in my heart, thank you for the friendship!

DEDICATION

献给我的父亲

Chapter 1

A NEW WAY TO TRADING: REDDIT WALLSTREETBETS

In this chapter, I will introduce (1) Reddit WallStreetBets (WSB) and its data, (2) Natural language processing to identify and construct relevant measurements, (3) the information content of Reddit WSB, and (4) the effects of Reddit WSB influencers.

1.1 GameStop and Reddit r/WallStreetBets

The GameStop stock event that unfolded at the beginning of 2021 represented a significant shift in the power dynamics of the stock market. Retail investors, particularly those active on social media platforms, demonstrated their ability to influence the market by driving the price of GameStop's stock up by 1,600% in a matter of days. This episode has been the subject of much debate regarding the role of retail investors and the potential risks associated with market speculation and manipulation. It also serves as the starting point of this dissertation, which examines the relationship between social media, retail trading, and the stock market.

The GameStop episode, also known as the GameStop short squeeze or the GameStop saga, refers to a series of events that took place in late January 2021 involving the stock GameStop, a struggling video game retailer. A short squeeze usually happens when the short sellers, investors who have bet against a particular stock, are forced to buy shares of the stock in order to limit their losses. This buying activity can cause the stock price to rise rapidly, exacerbating the losses of the short sellers and triggering more short sellers to close their positions. In other words, the buying behavior of the short sellers can create a cycle, where short sellers are forced to buy even more shares, and eventually close out their positions.

The GameStop short squeeze started when a group of individual investors on the subreddit r/wallstreetbets noticed that a number of hedge funds had taken out large short positions on GameStop's stock, specifically, according to data from a financial analytics firm, the short interest ratio for GameStop reached a high of 140% at the time. Keith Gill, a Reddit user known as "DeepF**kingValue" on the WallStreetBets subforum, who shared his long interest in GameStop and his investment return on r/WallStreetBets, inspiring many retail investors to build long positions in the stock despite the criticism from some prominent hedge funds such as Citron Research. On January 27, 2021, the stock reached a high of $347.51, forcing major hedge funds with short positions to close their positions at a loss. During this event, the stock price of GameStop increased to $500 at one point from its starting price of $17.25 in January 2021. Keith Gill's position, which he had shared on WallStreetBets, rose to a value of $48 million by January 27, 2021.

In recent years, Reddit r/WallStreetBets has gained widespread attention for its role in not only the GameStop episode but also for its influence on online investment communities. The subreddit was created in 2012 and has since grown to over 10 million members, as of early 2023. In many senses, Reddit r/WallStreetBets can be a good example and represent the generational shift in how younger retail investors use digital and social media to research and make informed decisions. In 2022, a NASDAQ survey stated, "Gen Z investors mostly use online discussion boards to gather stock trading information." It was clear that social media could be one of the information sources for younger investors, who are likely to take over the entire retail trading flow in a few decades. The role of the new generation of retail investors become increasingly important. To understand them, I start with one of their information sources, social media, specifically, Reddit r/WallStreetBets.

Before presenting the Reddit r/WallStreetBets data, it would be helpful to describe the Reddit forum for interested readers. Reddit is a social news aggregation, content rating, and discussion website that allows users to submit and share various forms of content, including links, text posts, images, and videos. The platform is home to numerous subreddits, which are individual discussion forums centered around particular topics or interests. WallStreetBets is

one such subreddit that focuses on stock and options trading. For the rest of the dissertation, I use WSB as short for r/WallStreetBets.

1.2 Reddit WSB Data

Reddit WSB data[1] used in this paper consists of two primary types of posts: submissions and comments. Submissions are the original content that users post to the site, while comments are responses to that content. Users can also reply to comments, creating threaded discussions that can span multiple levels. This structure enables robust and dynamic conversations on the site, as users can engage with content and each other in a variety of ways. See Appendix for snapshots of the Reddit WallStreetBets.

I retrieved the data from the PushShift.io Reddit application programming interface (API). It was designed and created by the moderator team of the subreddit r/datasets to help provide enhanced functionality and search capabilities for Reddit comments and submissions[2]. For my purposes, I only collect submissions and comments from WSB[3].

The WSB data spans from January 1st, 2018, to November 25th, 2021 and includes a comprehensive list of information, namely, the author of each submission/post, submission titles, descriptions, date/time, and number of comments. Thus, the dataset provides information on the online identity of the individual responsible for each submission or comment, as well as the date and time of their postings. Additionally, the number of comments received by a particular submission or comment can be used as a measure of its popularity. However, it is important to note that the dataset only includes publicly available information, and any content that was deleted or made private will not be reflected in the data.

Between 2018 and 2021, a total of more than 1.6 million submissions were recorded, encompassing a wide range of comment counts (from zero to thousands of comments). Analysis

---

[1]This work was facilitated though the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system at the University of Washington.

[2]For more information, see `https://reddit-api.readthedocs.io/en/latest/\#`.

[3]See `https://www.reddit.com/r/wallstreetbets/` for Wall Street Bets.

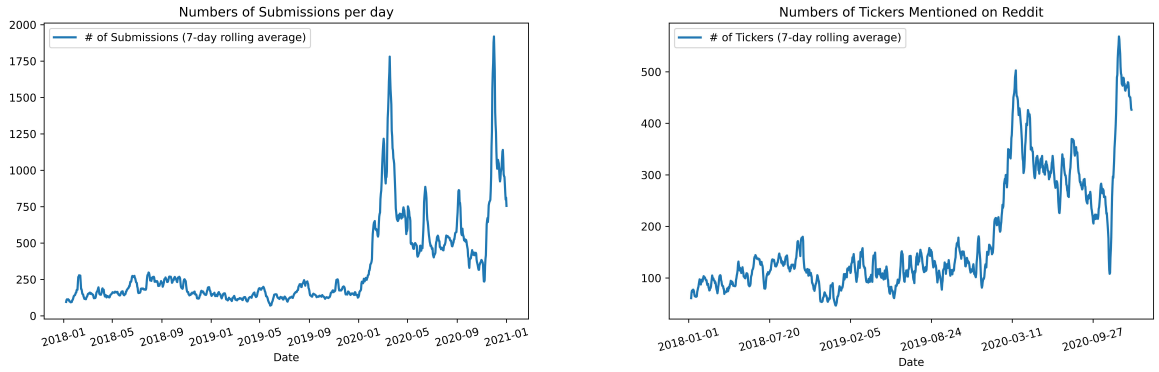


Figure 1.1: Time-series Plots on Reddit Activities

of the data on an annual basis reveals a notable upward trend in daily submission rates, with pronounced peaks observed at the onset of 2020 and 2021. Concurrent with the expansion of the r/WallStreetBets community, there has been a commensurate proliferation of stock tickers referenced in their discussions. Figure 1.1 shows the time-series plot describing the numbers of submissions (7-day rolling average) and the number of stock tickers (7-day rolling average).

1.2.1 Natural Language Processing

It should be noted that not all of the submissions retrieved in this study were related to stock trading. For example, on an average day in 2020, there were 654 submissions, of which only 202 pertained to stocks. To accurately identify and classify the relevant submissions, the following steps were taken to filter them and attach the corresponding stock tickers:

- Create a list containing the stock tickers that appeared in all submissions and comments. I use regular expressions to find all words or symbols followed by a money sign, as it is frequently used by social media users when discussing a particular firm's stock. For example, Reddit users could refer to Amazon.com, Inc.'s stocks as "$AMZN," and "AMZN" will be retrieved and appended to the list. One caveat is that Reddit users

could use the money sign with other terminologies. For example, "$YOLO" could mean the shares of AdvisorShares Pure Cannabis ETF or the popular jargon that WallStreetBets users frequently use as "You only live once." In fact, "$YOLO" would mostly mean the latter as it is part of the WSB culture of taking high risk for high returns. After carefully reviewing some of the submissions, I removed tickers that are deemed spurious from the list (See Appendix for more information).

- Use the list of stock tickers from the previous step to flag all submissions using regular expressions. This step only applies to submissions, not the comments. As a result, each submission will be mapped to one or more stock tickers, if applicable. For example, if the title or description of the submission contains "AMD" and "AAPL," the added stock ticker flag will be "AMD, AAPL." Usually, multiple tickers are mentioned in one submission. There are about 1/3 submissions that can be matched to stock tickers.

- After identifying all submissions, I proceeded to examine their comments as the final stage of data cleaning. Each comment underwent an identical procedure, wherein I augmented the corresponding submission tickers to the comment's tickers. Specifically, in instances where a submission contained tickers "AMD" and "AAPL," and a comment mentioned "TSLA," the resulting comprehensive list of tickers associated with that comment would be "AMD, AAPL, TSLA."

In the next section, I will present a series of measurements to gauge investors' attention and sentiment toward certain stocks on Reddit. It is crucial to note that the methodology described above would inevitably include tickers that have alternative meanings. However, without examining each and every submission/comment, I opted to keep most extracted tickers.

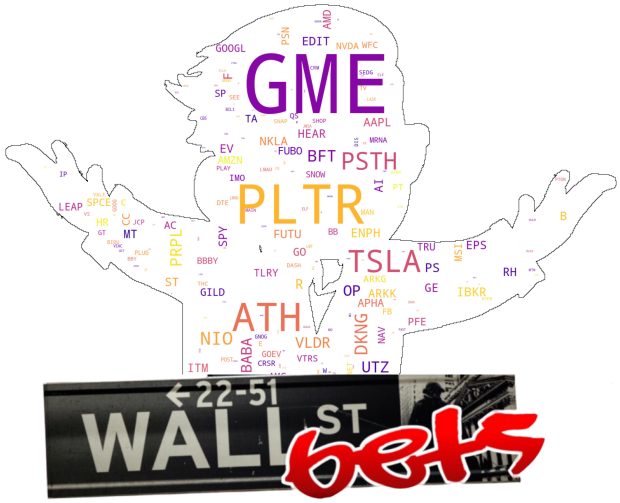


Figure 1.2: Reddit Ticker WordCloud in the shape of the famous WSB icon

1.2.2 Reddit Activity Measurements

Reddit Attention

The Reddit WSB Attention variable aims to measure Reddit users' attention towards specific stock tickers on a given day, and it is constructed as the total number of mentions for stock ticker i on the day t. In calculating the attention, I assume stock tickers mentioned in the submission weigh equally as those mentioned in the comments. For example, if there are 10 submissions and 300 comments about $AAPL, the daily Reddit attention for $AAPL will be 310.

It is worth noting that the moderator team of r/WallStreetBets would regularly create submissions entitled "Daily Discussion Thread" and "What Are Your Moves Tomorrow" on the previous day $(t-1)$. These submissions usually gain much attention among WSB users, generating thousands of stock-related comments. As a result, they serve as a potential representation of Reddit WSB users' perspectives on the stock market. On the contrary, combining stock mentions from all submissions and their corresponding comments on day t produces attention that includes both Ex-Ante and Ex-post content. Analysis of the timing

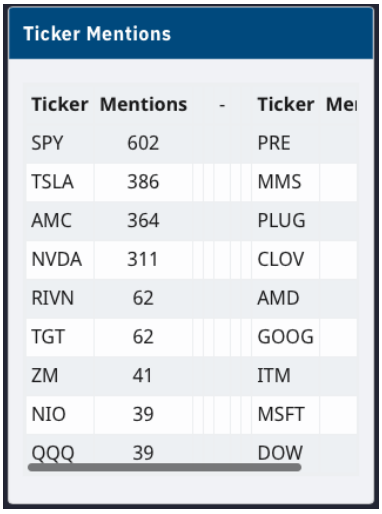


Figure 1.3: Stock Ticker Mentions On Reddit WSB

of the submissions in the collected data indicates that less than 20% of the submissions were made during market hours on a typical day. This implies that the Reddit WSB attention variable constructed from all submissions and comments is predominantly composed of Ex-post information. Both measurements, namely those based on the "Daily Discussion" and those based on all WSB content, are employed throughout the paper, with greater emphasis placed on the "Daily Discussion" metric. More information about the two variables can be found in the table 1.1. As I wrapping up the studies, Reddit WSB had created a similar metric and displayed it on the main page of the subreddit[4], see Figure 1.3. What's more, there are multiple trading platforms that have also collected the same metric, and presented it as one of the important market information (see Appendix).

Reddit Sentiment

The Reddit WSB sentiment variable aims to measure the WSB users' attitudes toward individual stocks. I employed the lexical methods, specifically the Laughran-Mcdonald dictio-

[4]Screenshoted on Feb. 28th, 2023.

nary and Valence Aware Dictionary and sEntiment Reasoner (VADER, to extract sentiment scores from textual data. The lexical method typically starts with a pre-existing dictionary of sentiment words, along with the corresponding sentiment scores. The sentiment scores are then assigned to texts that contain the identified sentiment words. The effectiveness of this method is reliant on the comprehensiveness of the sentiment dictionary and may not yield scores for all texts. In contrast, one can choose to use the machine learning method, which requires a large training dataset with predefined sentiment scores. Creating a training dataset can be labor-intensive and potentially error-prone due to differing interpretations of sentiment by different individuals. In the subsequent sections, I will discuss the two sentiment dictionaries utilized in my analysis.

Laughran-McDonald Dictionary The Loughran-McDonald master dictionary with sentiment word lists is based on Loughran and Mcdonald(2011). It has been updated over the years to account for new sentiment words[5]. The dictionary combines words from the EDGAR 10-X fillings with base words from the English dictionary. Only those words that appear at least 100 times in the 10-X fillings and that can be identified as actual words are added to the dictionary. The LM dictionary worked well with financial text and has been applied to social media in previous studies. For example, Sanford (2022) used the Loughran-McDonald dictionary to create daily sentiment indices for Twitter data. The dictionary comprises seven categories, namely negative, positive, uncertainty, litigious, strong modal, weak modal, and constraining. I used only the negative and positive categories of sentiment in this paper. In total, the dictionary contains 2,709 words that are associated with either positive or negative sentiment. For instance, the term "GREAT" represents a positive sentiment (+1), while "TERRIBLE" indicates a negative sentiment (-1).

VADER The second lexical method used is VADER (Valence Aware Dictionary and sEntiment Reasoner). It is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. VADER could be an excellent substitute

[5]The Loughran-McDonald master dictionary with sentiment word lists (updated in January 2022) was downloaded from the Software Repository for Accounting and Finance, University of Notre Dame.

for the LM dictionary because it can analyze not only the colloquialism (emoji like a grinning face, acronyms like "LOL," slang like "friggin") but also the punctuations and word-shape that signal increased sentiment intensity ("good!" v.s. "good!!!," "good" v.s. "GOOD"). Moreover, it can consider the context before assigning a sentiment score to the word. For example, the term "catch" has negative sentiment in "At first glance, the contract looks good, but there's a catch" but is neutral in "The fisherman plans to sell his catch at the market." These characteristics allow VADER to give a sentiment score to almost all text. However, one caveat is that VADER only works better with shorter texts, and tends to assign neutral scores to longer texts. VADER gives four categories of sentiment, positive, negative, neutral, and compound scores. The compound score is computed by summing the sentiment scores of each word and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). In this study, I used compound score as the VADER sentiment measurement for submissions/comments[6]. Interested readers can go to the VADER GitHub website for more information[7].

The WSB sentiment for each submission/comment is constructed in the following way: (1) utilizing natural language processing to divide submissions/comments into sentences, (2) applying either LM dictionary or VADER to identify the sentiment scores, and (3) WSB sentiment for a particular stock on a given day is calculated as the average of the sentiment scores form all associated submissions/comments.

Sample description and summary statistics

Before proceeding to the next part of my study, it is useful to summarize the Reddit data. From the Reddit-related data, I created a panel dataset. For each ticker-date pair, the data consists of several variables, including the date, ticker, Reddit WSB attention (based on all

[6]Typical thresholds for classifying sentences as either positive, neutral, or negative are: positive sentiment when the compound score is greater or equal to 0.05; neutral sentiment when the compound score is in between -0.05 and 0.05, and negative sentiment when the compound score is smaller or equal to -0.05.

[7]VADER Sentiment Analysis, website: `https://github.com/cjhutto/vaderSentiment`

content or "Daily Discussion" submissions), and Reddit WSB sentiment (based on either LM dictionary or VADER). All Reddit WSB attention variables are calculated as the total number of mentions for stocks on a day. All WSB sentiment variables are constructed as the average sentiment scores for stocks on a day. In the subsequent section, I examine the information content of the Reddit data. In order to understand the information content of Reddit WSB, stock daily return data from CRSP, and Key developments from Capital IQ were selected into my data. The key development database provides structured summaries of material news and events that may impact the market value of securities. It records event types that include dividends, M&A, buybacks, public offerings, management changes, debt defaults, and more. The following table 1.1 reports the summary statistics of key variables used in my analysis.

1.3 Understanding the Information Content

Prior studies have demonstrated the ability of Reddit WSB to influence certain stocks within short timeframes. For instance, Long et al. (2023) examined the role played by WSB in the GameStop event and established the relationship between WSB sentiments on 5-,10-, and 30-min GME returns. Anand and Pathak (2022) similarly demonstrated that Reddit tone impacts GME volatility, spread, and volume for up to 10 mins in advance. Furthermore, Cioroianu et al. (2021) utilized a Twitter-based sentiment index and found the impact of social media created a unique form of information asymmetry and market euphoria, leading the investors to focus on short-term profitability of the firm around blockchain-related announcements. With the ample research on Reddit's impact on the market, one has to wonder about the longer effect that Reddit has on the stock market, and whether it extends to a broader range of stock tickers. In the next part, I will examine whether (1) Reddit activities can predict stock return in a 10-day timeframe, and (2) Reddit WSB responds to the key developments that happened to the firms. This analysis can help understand the information content of Reddit WSB, and importantly, whether WSB users respond to firms' fundamental information. Key development events are considered instances where firms' fun-

Table 1.1: This table shows the summary statistics of Reddit-related, Key developments, and stock return data. The time period of the data is January 1st, 2018 to November 25th, 2021. In panel A, I present all Reddit-related variables. *Reddit_Attn_DD* is the total number of stock mentions for a particular stock on a given day, based on the Reddit submissions entitled "Daily Discussion Thread" and "What Are Your Moves Tomorrow" and their comments. *Reddit_Attn_All* is the same measurement based on all submissions/comments on Reddit for a stock ticker on a given day. *Reddit_Sent_LM* is the average sentiment score based on the Laughran-McDonald Dictionary. *Reddit_Sent_VADER* is the average compound sentiment score using VADER. In panel B, *KeyDev* is an indicator variable of key developments for a stock on a given day, collected from the Capital IQ Key Developments database. It equals one when there are key developments and zero when there are not. *NumOfKeyDev* counts the number of key developments for stocks on a day. *Return* is the daily stock return from CRSP.

Panel A

	Count	Mean	Std Dev	Min	p50	max
Reddit_Attn_DD	3,282,590	0.709	15.108	0	0	6,363
Reddit_Attn_All	4,526,172	0.677	65.646	0	0	61,595
Reddit_Sent_LM	65,873	-0.554	2.673	-117	-1	39
Reddit_Sent_VADER	213,539	0.127	0.348	-0.99	0.093	1

Panel B: Others

	Count	Mean	Std Dev	Min	p50	max
KeyDev	3,282,590	0.058	0.234	0	0	1
NumOfKeyDev	3,282,590	0.085	0.399	0	0	12
Return	8,433,609	0.0005	0.0410	-0.9286	0	12.3652

damental information is exchanged. The most widely reported retail trading around Reddit WSB, namely $GME, $BBB, and $TSLA, is tightly connected to so-called Meme stocks, and when expressing trading opinions, WSB users often say "I just like the stock", instead of providing a fundamentals-related reason for the trades. This section aims to provide a more clear picture as to why WSB users "just like the stock".

1.3.1 Reddit's return predictability

In this section, I present the predictive regressions of the return on a 10-day horizon as the following:

$$Ret_{i,t+d} = b_0 + b_1 Reddit\ Measurements_{i,t} + b_3 Controls + FE_t + e_{i,t} \tag{1.1}$$

For stock ticker i on the day t, $Reddit\ Measurements_{i,t}$ could be (1) Reddit Attention, measured by daily numbers of stock mentioned (standardized cross-sectionally)[8], or (2) Reddit Sentiment, which is an indicator variable, constructed based on VADER sentiment. The indicator equals 1 if the VADER sentiment is positive, and 0 otherwise. Note that the missing values were dropped from the regression, and I did not fill in zeros when there is missing values. If one were to fill all the missing values with zero, essentially, the assumption made is that Reddit WSB holds a non-positive sentiment toward the stock on a given day. This can be a fairly big assumption to make, and it was unclear what bias this assumption would introduce into the analysis. The disadvantage of not filling in missing values is my analysis would automatically have fewer data points available. I fitted this regression with a date fixed effect to control for date-variant factors, and to account for the time-series correlation for a given firm, the standard errors are clustered at the firm level.

The results are reported in table 1.2. In panel A, 10 predictive regressions were conducted, with Reddit Attention and controls as the independent variables. Reddit Attention appears to have a persistently negative effect on the return, with the coefficients ranging from -

[8]Based on submissions titled "Daily Discussion" or "What Are Your Moves", and their comments.

0.595 (t-stats -3.206) to -0.196 (t-stats -1.690). In a 10-day window, the return is negatively related to the Reddit Attention for 5 out of the 10 days. In panel B, When using another measurement - the WSB sentiment, positive sentiment is shown to be significantly associated with a higher return for 9 out of the 10 predictive regressions. The coefficients of the Reddit sentiment range from 1.328 (t-stats 2.765) to 3.497 (t-stats 3.625). Most interestingly, when interacting the WSB attention with sentiment measurements, the overall effect is positive for, again 9 out of the 10 days, however, in the regressions for day $t+3$ and $t+6$, the interaction term between sentiment and attention yield negative coefficients, -1.072 (t-stats -2.381) and -0.572 (t-stats -3.061), respectively. A concentrated positive opinion toward stocks can be hurtful to the returns.

In summary, this analysis finds that Reddit has a consistent and strong predictive power on stock returns, as indicated by multiple statistically significant coefficients. While the GameStop event was unique and may not recur, Reddit remains a valuable predictor of 2-week stock returns. This suggests the information content on Reddit is not entirely noise, and can potentially contain useful information about the firm value. However, I also find that concentrated positive sentiment towards certain stocks on Reddit could be harmful to the return. This is evidenced by the negative interaction term between WSB attention and sentiment. This suggests that it's always useful to look at various WSB activities as they can differ when it comes to stock return predictability, among other things.

Table 1.2: This table details the return predictive power of Reddit activities. Panel A, B, and C show regressions of returns on Reddit activities. The dependent variable is the daily stock return from day t to $t+10$. The independent variables, *Reddit_Attn_DD (i,t)* is the total number of mentions for stock i on day t, based on the "Daily Discussion" submissions and comments. *Reddit_Sentiment(i,t)* is the indicator of Reddit sentiment constructed using VADER. It equals to 1 if the VADER sentiment is positive, and equals to 0 otherwise. There are two firm-level controls, namely *ln(market cap)*, and *ln(bm).ln(market cap)* is the natural logarithm of firm's market capitalization. *ln(bm)* is the natural logarithm of firm's book-to-market ratio. To account for the impact of historical returns, I added *ret (t-1)*, *ret (d2-d20)*, and *ret (d21-d120)*, which are stock return on day $t-1$, cumulative stock return from day $t-20$ to day $t-1$, and cumulative stock return from day $t-120$ to day $t-21$, respectively. to The regressions are fitted using panel regression with date fixed effect and clustered standard error of date. Controls are the same in all panels. Robust t-statistics are reported in parentheses. ***,**,* indicate the significance level of 1%, 5%, and 10%, respectively. The more relevant coefficients are in bold.

Panel A: Regression between Return and Reddit Momentum

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
VARIABLES	*Return (t)*	*Return (t+1)*	*Return (t+2)*	*Return (t+3)*	*Return (t+4)*	*Return (t+5)*	*Return (t+6)*	*Return (t+7)*	*Return (t+8)*	*Return (t+9)*	*Return (t+10)*
Reddit_Attn_DD (t)	**-0.196***	**-0.366***	**-0.375****	-0.384	**-0.301***	-0.209	**-0.286****	**-0.595*****	-0.264	0.0408	**-0.434***
	(-1.690)	(-1.851)	(-1.968)	(-1.385)	(-1.925)	(-1.498)	(-2.030)	(-3.206)	(-1.440)	(0.587)	(-1.935)
Observations	646,239	479,027	349,655	335,001	346,092	501,108	618,816	596,258	458,220	344,785	345,257
R-squared	0.410	0.186	0.001	0.057	0.124	0.084	0.203	0.127	0.152	0.082	0.216
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 1.3: This is a continuation of the previous table.

Panel B: Regression between Return and Reddit Sentiment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
VARIABLES	*Return (t)*	*Return (t+1)*	*Return (t+2)*	*Return (t+3)*	*Return (t+4)*	*Return (t+5)*	*Return (t+6)*	*Return (t+7)*	*Return (t+8)*	*Return (t+9)*	*Return (t+10)*
Reddit_Sentiment (i,t)	**1.382*** **	**2.145** **	**3.497*** **	**3.344*** **	**1.813***	**2.429*** **	**3.159*** **	**3.342*** **	**2.628*** **	1.134	**2.608*** **
	(2.765)	(2.444)	(3.625)	(2.626)	(1.956)	(3.235)	(3.979)	(4.106)	(3.048)	(1.394)	(2.706)
Observations	885,998	714,371	585,238	569,964	582,702	738,587	856,107	830,530	692,035	579,123	580,127
R-squared	0.441	0.211	0.001	0.057	0.105	0.078	0.191	0.118	0.138	0.082	0.216
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Panel C: regression between Return and Reddit Attention & Sentiment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
VARIABLES	*Return (t)*	*Return (t+1)*	*Return (t+2)*	*Return (t+3)*	*Return (t+4)*	*Return (t+5)*	*Return (t+6)*	*Return (t+7)*	*Return (t+8)*	*Return (t+9)*	*Return (t+10)*
Reddit_Sentiment (i,t)	**1.584*** **	**2.698*** **	**3.812*** **	**3.708** **	**2.245** **	**2.351*** **	**3.384*** **	**3.649*** **	**2.885*** **	1.007	**3.252*** **
	(2.821)	(2.801)	(3.065)	(2.485)	(2.310)	(2.680)	(3.687)	(3.916)	(2.874)	(1.097)	(2.613)
Reddit_Attn_DD (i,t)	0.0172	-0.288	-0.298	**0.662** **	0.350	-0.468	0.269	-0.396	-0.316	-0.0418	-0.818
	(0.115)	(-1.051)	(-0.952)	(1.996)	(0.277)	(-1.381)	(1.496)	(-1.213)	(-0.921)	(-0.128)	(-1.191)
Interaction Term	-0.219	-0.0794	-0.0782	**-1.072*** **	-0.671	0.268	**-0.572*** **	-0.204	0.0545	0.0855	0.393
(-1.304)	(-0.357)	(-0.237)	(-3.456)	(-0.523)	(0.804)	(-3.061)	(-0.671)	(0.203)	(0.271)	(0.580)	
Observations	646,239	479,027	349,655	335,001	346,092	501,108	618,816	596,258	458,220	344,785	345,257
R-squared	0.410	0.186	0.001	0.057	0.124	0.084	0.203	0.127	0.152	0.082	0.216
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

1.3.2 Reddit's Response to Key Developments

The second channel of understanding Reddit's information content is through its responsiveness to firms' key development events. It is unclear whether Reddit activities should respond to the key developments based on the previous studies. Professional and institutional investors closely monitor firms' key developments, engage in thorough analysis of analysts' reports and utilize high computational power to make investment decisions. Conversely, users of the WSB appear to be driven by emotions and trade "meme stocks", as evidenced by statements such as "I just like the stock." After the GameStop event, Keith Gill concluded his testimony before the U.S. House Committee on Financial Services by affirming his continued commitment to his position in GME with the declaration "In short, I like the stock." Stocks like GameStop are often referred to as "meme stocks" in many studies. Chacon et al. (2022) found that the "meme stocks" portfolio is not profitable on a risk-adjusted basis, and holding periods ranging from one day to one year failed to produce alpha. Retail sentiment, as suggested by Kumar and Lee (2006), is not a simple artifact of news events typically associated with changes in stock fundamentals, such as macroeconomic news or analyst earnings forecast revisions. Moreover, retail traders appear to be indifferent to earnings news in the presence of media sentiment, as suggested by Cahill et al. (2017).

On the flip side, as an experience Reddit WSB user, it is not uncommon to come across submissions/comments citing analyst's reports, earning calls, and other major events for some stocks. In fact, WSB has a flair setup for such posts, DD, which stands for "Due Diligence". In the context of WSB, DD typically refers to research and analysis of a specific stock or company in order to determine whether it is a good investment opportunity. This can involve reviewing financial statements, analyzing industry trends, and assessing the company's competitive position, among other factors. DD is often emphasized on WSB as a way for users to share information and insights that may help others make informed investment decisions. Bradley et al. (2021) discovered that over the 2018-2020 sample, DD recommendations on Reddit WSB are significant predictors of one-month ahead returns, earnings forecast

revisions, and earnings surprises. In addition, user comments are incrementally useful for predicting returns, and small retail trade informativeness increases following DD reports. However, this result reversed in 2021 after the GME event happened, suggesting the event and its subsequent effects on the subreddit deteriorated its usefulness for smaller investors.

In summary, it remains unclear whether Reddit WSB reacts to the fundamentals, proxied by the key developments in this study. The following model is used to examine whether WSB Redditors respond to the fundamental information:

$$Reddit\ Activites_{i,t} = b_0 + b_1 NumofKeyDev_{i,t} + FE_{i,t} + e_{i,t} \tag{1.2}$$

In this regression, the independent variable is Reddit activities, and it can be measured by Reddit Attention or Reddit Sentiment. The independent variable *NumofKeyDev* is the total numbers of key developments that happened for firm i on date t. To account for firm- and time-specific variables, the firm and date fixed effect is added into the model. Clustered standard error for date is used. If the coefficient, b_1, is positive and significant, then Reddit users are shown to respond to the Key developments.

In table 1.4, I present the result. On average, WSB subredditors do respond to the key developments, and as the numbers of key developments increases by one, the Reddit attention observed in "Daily Discussion" posts increases 0.559 (t-stats 12.17), which translates to 0.559 increase in the number of stock mentions. The Reddit sentiment, measured by sentiment words from the Laughran-McDonald dictionary, established a negative and significant relationship with *numofkeydev*. When the numbers of key developments increases by one, the Reddit sentiment decreases by 0.0668 (t-stats -3.713) on average. Therefore, despite the anecdotal evidence that suggests WSB users do not trade on fundamentals, they appear to respond to firms' key development events. The regressions on the other Reddit metrics can be found in the Appendix.

When referring to the key developments of the firms, there can be many types of events, namely Analyst/Investor Day, Announcement of Operating Results, Announcements of Earnings, Annual General Meeting, etc. The previous result suggests Redditors respond to the

Table 1.4: This table shows the relationship between various Reddit WSB activites and Key Developments. The dependent variable, *Reddit_Attn_DD* is total numbers of mentions for firm i on day t, based on "Daily Discussion" submissions and comments. *Reddit_Sent_LM* is the Reddit WSB sentiment score for firm i on day t, using Laughran-McDonald Dictionary. The independent variable, $numofkeydev$ is the total number of key development events that happened for firm i on day t. The regression is fitted with firm and date fixed effects. Standard error is clustered by date. Robust t-statistics are reported in parentheses. ***,**,* indicate the significance level of 1%, 5%, and 10%, respectively.

	(1)	(3)
VARIABLES	*Reddit_Attn_DD*	*Reddit_Sent_LM*
numofkeydev	0.559***	-0.0668***
	(12.17)	(-3.713)
Observations	3,282,590	53,330
R-squared	0.111	0.147
Date FE	Yes	Yes
Ticker FE	Yes	Yes

events on average, however, it was not clear which types of key events affect Reddit activities the most. This is an important question to ask because the information that stimulates Reddit attention/sentiment can help categorize what type of traders they are. As Barber et al. (2021) noted, the retail traders can have adverse reactions to a positive or neutral news. For instance, Autora Cannabis Inc ($ACB) had a 1-to-12 reverse stock split. The share price had dropped to $0.67, which required the reverse split to keep the stock listed. This merely mechanical event had led to some retail investors abandoning the stock and never coming back despite the consecutive high returns in the following days.

For the next set of results, I identified the top 10 key events for Reddit ticker universe. From 2018 to 2021, I counted the occurrences of each type of events that happened, and ranked them to obtain the top 10 key events. They are Conference, Company Conference Presentations, Earning Announcement, Product-related Announcements, Earning Call Time, Earning Release Date, Executive Board Changes, Client Announcement, Buyback Tranche Update, and Corporate Guidance(See Appendix for the description of the top 10 key events).

The next two tables report the results. Overall, Reddit Attention is related to most top 10 key events, with the exception of Company Conference Presentations (when a company participates in a "conference call" conducted by third parties), Earning Call Time (information about when the earning call will be conducted), and Earning Release Date (information about when the earning information will be released). Particularly, Reddit attention increases by 3.601 (t-stats 10.67) when there is an earning announcement, and 3.816 (t-stat 10.09) when there is a corporate guidance event. The coefficients of Client Announcements and Executive Board Changes are also significant and positive. However, when looking at the Reddit sentiment, overall WSB users expressed negative sentiment around Conferences, Earning Announcements, Buyback Tranche Update, and Corporate Guidance. For instance, when there is a corporate guidance, the sentiment drops by 0.387 (t-stats -3.175). What's more, on average, there is a -0.267(t-stats -2.828) sentiment change when there is an earning announcement.

Table 1.5: This table shows the relationship between Reddit Attention and different types of Key Developments. The dependent variable, *Reddit_Attn_DD* is total numbers of mentions for firm i on day t, based on "Daily Discussion" submissions and comments. The independent variable, $numofkeydev$ is the total number of key development events that happened for firm i on day t. The regression is fitted with firm and date fixed effects. Standard error is clustered by date. Robust t-statistics are reported in parentheses. ***,**,* indicate the significance level of 1%, 5%, and 10%, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	*Reddit_Attn_DD*	*Reddit_Attn_DD*	*Reddit_Attn_DD*	*Reddit_Attn_DD*	*Reddit_Attn_DD*	*Reddit_Attn_DD*	*Reddit_Attn_DD*	*Reddit_Attn_DD*	*Reddit_Attn_DD*	*Reddit_Attn_DD*
Conferences	0.240**									
	(2.021)									
Company_Conference_Presentations		0.124								
		(1.336)								
Earning_Announcement			3.601***							
			(10.67)							
Product_Related_Announcements				0.994***						
				(4.828)						
Earning_Call_Time					0.122					
					(1.397)					
Earning_Release_Date						0.137				
						(1.592)				
Executive_Board_Changes							0.539***			
							(3.586)			
Client_Announcements								1.150***		
								(4.991)		
Buyback_Tranche_Update									2.684***	
									(8.896)	
Corporate_Guidance										3.816***
										(10.09)
Observations	3,282,590	3,282,590	3,282,590	3,282,590	3,282,590	3,282,590	3,282,590	3,282,590	3,282,590	3,282,590
R-squared	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 1.6: This table shows the relationship between Reddit Attention and different types of Key Developments. The dependent variable, *Reddit_Sent_LM* is the WSB sentiment, built using Laughran-McDonald Dictionary. The independent variable, $numofkeydev$ is the total number of key development events that happened for firm i on day t. The regression is fitted with firm and date fixed effects. Standard error is clustered by date. Robust t-statistics are reported in parentheses. ***,**,* indicate the significance level of 1%, 5%, and 10%, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	*Reddit_Attn_LM*	*Reddit_Attn_LM*	*Reddit_Attn_LM*	*Reddit_Attn_LM*	*Reddit_Attn_LM*	*Reddit_Attn_LM*	*Reddit_Attn_LM*	*Reddit_Attn_LM*	*Reddit_Attn_LM*	*Reddit_Attn_LM*
Conferences	-0.183**									
	(-1.981)									
Company_Conference_Presentations		-0.108								
		(-1.519)								
Earning_Announcement			-0.267***							
			(-2.828)							
Product_Related_Announcements				-0.102						
				(-0.864)						
Earning_Calls					-0.00290					
					(-0.0217)					
Earning_Release_Date						0.119				
						(0.891)				
Executive_Board_Changes							-0.201			
							(-1.589)			
Client_Announcements								-0.0470		
								(-0.395)		
Buyback_Tranche_Update									-0.249**	
									(-2.240)	
Corporate_Guidance										-0.387***
										(-3.175)
Observations	53,330	53,330	53,330	53,330	53,330	53,330	53,330	53,330	53,330	53,330
R-squared	0.147	0.147	0.147	0.147	0.146	0.146	0.147	0.146	0.147	0.147
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

1.4 Reddit Influencers

As emphasized by Shiller (1989) page 7, "[I]nvesting ... is a social activity. Investors spend a substantial part of their leisure time discussing investments, reading about investments, or gossiping about others' successes or failures in investing." Since day one, communication has influenced equity trading. In the 17th century, most brokers and investors did their business in the various coffee shops around London. Debt issues and shares for sale were written up and posted on the shops' doors or mailed as a newsletter.

Reddit WSB is a modern example of the social aspect of trading. Redditors talk about various stocks and derivatives on the platform, with the intention of gathering information, sharing gain/loss, and searching for investment opportunities. There are comprehensive theoretical studies that focus on the social network effects in the financial market. For example, a standard DeGroot (1974) model describes how a group with multiple individuals might reach an agreement on a common subjective probability distribution. DeMarzo et al. (2003) introduces a model where individuals are subject to persuasion bias, in other words, they fail to account for possible repetition in the information they receive. A recent theoretical study, Pedersen (2021) presents a model of how investment ideas can propagate through a social network and affect market behavior and prices. The model has several important implications for how social networks affect opinions and asset prices. Importantly, it emphasizes the role that hard-headed fanatics play in influencing the expected stock demand. In the paper, the author explained how the model can fit into explaining the GameStop episode. It also provides a channel for the idiosyncratic changes of opinion spur to others and eventually affect the community. There is also a growing body of empirical literature that focuses on the social network perspective of Reddit. For instance, Hu et al. (2021) examines the relationship between Reddit connectedness and stock returns. Yousaf et al. (2023) analyzes the quantile connectedness between meme stocks, meme tokens, and traditional financial assets. However, the effects that influencers, or particularly, Reddit WSB influencers have on the general public are yet to be examined over a longer time period. Numerous studies have in-

vestigated Reddit WSB top Redditors' influencers around the GameStop episode, and Keith Gill, as the most important Redditor in the event, is frequently mentioned in the studies. As GameStop is unique in many different ways, I think it is important and interesting to quantify the effects of Reddit's influencers over a longer period of time and expand my analysis to more influential Redditors.

In the next section, I will briefly describe the main results of a recently developed theoretical framework in Pedersen (2021). As mentioned before, it provides a good explanation of the GameStop episode and some important empirical implications that I intend to test.

1.4.1 Theory Framework

The Pedersen (2021) model introduces rational agents and financial markets into an otherwise standard DeGroot (1974) model.

Model Setup In the economy, The asset value comprises a publicly observed random walk and an unobserved random variable that investors are trying to learn. There are four types of investors, naive, fanatic, rational long-term, and rational short-term investors. All of them will engage in multiple rounds of talks to gather information. Each naive investor selectively follows a subset of people that he views as most informative or most entertaining and updates his information. Note that naive investors are subject to persuasion bias, which is they will repeatedly update their information when they hear about the same information for multiple rounds. Rational investors listen to everyone and form their opinion in the first round. Rational investors will not change their opinions and become stubborn. Fanatics only listen to themselves and will become stubborn.

Social Networks After a while, the dynamics boil down to how hardheaded (they can be fanatics or rational) interacts with the naive investors. As the naive investors selectively listen to a group of other investors, in the long run, each naive agent's view is a convex combination of the views of fanatics and rational agents. In this case, investors never reach a consensus, but they will have an average opinion. To get to the average opinion, each investor is assigned with an "influencer value", which is how much the investors influence

others. Eventually, the expected stock price and return of the entire social network will have two components, the hard-headed rational and fanatic opinions, weighted by their respective "influencer value".

Empirical Implications Idiosyncratic attention or sentiment shocks can lead to aggregate fluctuation in retail attention/sentiment. Without spelling out all the mathematics, I will summarize the main implication as "Reddit WSB's opinion" is the hard-headed opinions, weighted by hard-headed redditors' social influences in the long run.

Hypothesis and Approach Based on the theoretical framework described above, I have formed the following hypotheses and approaches to test each one of them: (1) Influencers' attention and sentiment should have some predictive power to WSB's aggregated attention and sentiment, either on the same day, or lagged, (2) If the influencers are mostly hard-headed fanatics (they only listen to themselves and do not update their belief), then their attention/sentiment should not be related to the firms' key development events. On the flip side, if their activities are more responsive to the key development events than a typical Redditor, it could suggest they listen to the fundamental information, and (3) If they are hard-headed and rational, then their sentiment/attention should establish a stronger relationship with stock returns compared to the general public on Reddit WSB.

In the next section, I will introduce some stylized facts about WSB influencers, followed by regression results for the above-mentioned hypotheses.

1.4.2 Stylized Facts about Influencers

I define the term "WSB influencers" as a cohort of Redditors active on the r/wallstreetbets subreddit who received a high volume of comments between 2018 and 2021. To identify this group, submissions with at least one comment were selected, and the total number of comments received by each Redditor was aggregated from January 2018 through November 2021. Then the top 10% of the cohort is identified as WSB influencers[9]. Notably, comments

[9]The choice of 10% is mainly based on the availability of data, and the sentiment data for the top 10% are still limited. Robustness checks for 1% show no major changes in the results.

that received high attention were not included. I focus on submissions only, which are conversations initiated by potential influencers. The influencers' sentiment scores are replaced with zeros if missing. In other words, intrinsically, I assume influencers are aware of the entire stock ticker universe, and the lack of mentioning indicates neutral sentiment. What's more, It is not uncommon for individuals to have multiple usernames or "aliases." As such, when defining influencers within the platform, it is important to note that the focus is on individuals' online identities rather than their real-world identities.

In the rest of the subsection, I will report some stylized facts about the influencers. Many papers documented that after the GameStop event in January 2021, there is a structural change on the Reddit WSB in terms of how users interact and the informativeness of certain types of posts. For instance, Bradley et al. (2021) examines the market consequences of due diligence (DD) reports on Reddit's Wallstreetbets (WSB) platform. Over the 2018-2020 sample, they find that DD recommendations are significant predictors of one-month ahead returns, earnings forecast revisions, and earnings surprises. However, all of these benefits reverse in the first half of 2021. In my analysis, I covered the period before GameStop and after GameStop to understand the average effect of Reddit influencers. In table 1.7, I report the top 12 influencers. Although "u/Deep*******Value", or Keith Gill, has been shown to be one of the key players in the GameStop episode, he only ranked the 12th when I expand the time horizon to 4 years. Notably, this list of influencers uses data till 2021, and some of the influencers listed are no longer active as of March 2023 (for example, "u/Deep*******Value" and "u/stormwillpass"). Table 1.8 reports the popular tickers and sentiment words among Reddit influencers. The influencers' tickers largely overlap with the entire WSB ticker universe. Mechanically, there are fewer tickers mentioned by influencers compared to the general public on WSB, however, the most popular tickers remain similar, including "$GME","$BB", and "$AMZN". The top 5 sentiment words that can be matched with the Laughran-McDonald dictionary are "Good","Best","Great","Better", and "Gains".

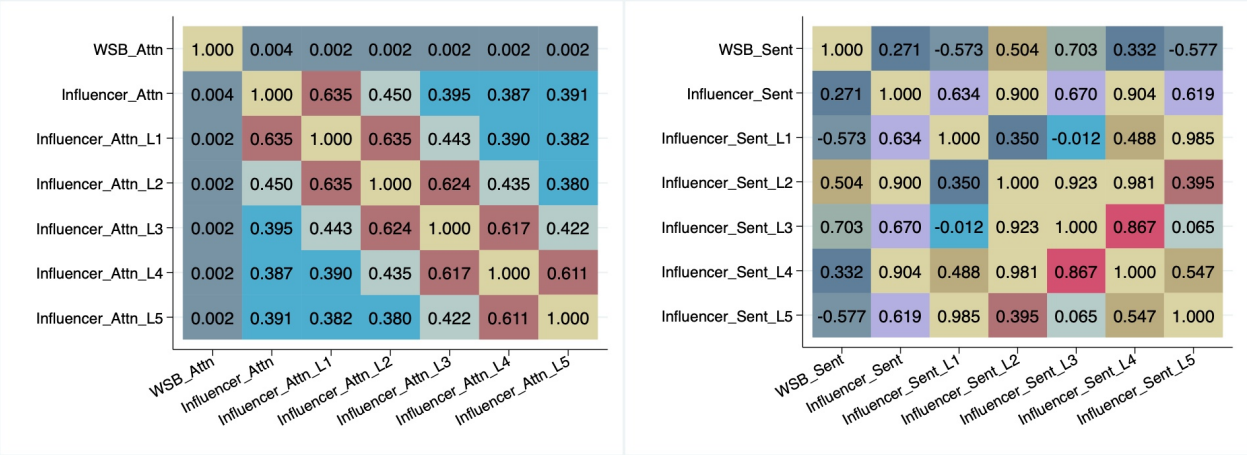
Table 1.7: This table lists the top 12 influencers on Reddit WSB (top 10 users who are not moderators). They are ranked by the sum of number of comments received from January 2018 to November 2021. Reddit WSB users' names always start with a "u/". The top two Redditors are on the WSB moderator teams and responsible for posting the "daily discussion" and "What are your moves tomorrow" submissions, thus they are ranked the top two.

Redditor	Comments	Description
u/AutoModerator	12,788,391	Moderator
u/OPINION_IS_UNPOPULAR	9,553,521	Moderator
u/grebfar	410,105	
u/stormwillpass	407,264	(Previous) Moderator
u/MotorizedDoucheCanoe	400,132	
u/wallstreetboyfriend	350,965	
u/zjz	342,479	
u/VisualMod	288,697	
u/theycallmeryan	206,879	
u/GoBeaversOSU	173,102	
u/Stylux	169,227	
*u/Deep*******Value*	116,563	Keith Gill (GME event)

Table 1.8: This table contains popular tickers and Laughran-McDonald sentiment words among WSB influencers from Juanary 2018 to November 2021. Influcers are defined as a cohort of Redditors who received the top 10 percentile numbers of comments.

Panel A: Popular Tickers among Influencers				
$GME	$BB	$AMZN	$FB	$NOK
$SPY	$EDIT	$ET	$MU	$NVDA
$TSLA	$AAPL	$NIO	$ITM	$ATH
$AMC	$PLTR	$MSFT	$SPCE	$BABA
$AMD	$EV	$RH		

Panel B: Popular Laughran-McDonald Sentiment among Influencers				
GOOD	*BETTER*	*BAD*	*LOSS*	*LOST*
BEST	*GAINS*	*CLOSED*	*GAIN*	*BREAK*
GREAT	*STRONG*	*ABLE*	*POSITIVE*	*LATE*
VOLATILITY	*OPPORTUNITY*	*LOSE*		



(a) Attention **(b)** Sentiment

Figure 1.4: Correlation between WSB Activities and Influencers' Activities (5-day lags)

1.4.3 Influencers: influencing or not

Hypothesis 1: Influencers' activities can predict the WSB activities

Theoretically, influencers can affect other people's opinions because they have larger audiences and their opinions are hard-headed. On a very high level, this means influencers' activities can help predict the general Redditor's activities. Figure 1.4 provides a sneak peak of the relationships.

To test this hypothesis, I used the following regression:

$$Reddit\ Activites_{i,t} = b_0 + b_1 Influencers\ Activities_{i,t} + Controls + e_{i,t} \tag{1.3}$$

Where the *Reddit Activites* is the Reddit attention or sentiment for stock i on day t, and the *Influencers Activities* are the corresponding activity from influencers. The model also includes two control variables, $ret(t-1)$ and $vol(t-1)$, which are the stock return and the stock volatility on day $t-1$, respectively. They are included because many studies have mentioned retail traders seek short-term attention, namely, stocks with high return/volatility on $t-1$. For example, Ozik et al. (2021) established that retail traders seek stocks that have

previously gained attention. The ticker and date fixed effects are also considered. All models shown in the results table (Table 1.9 and 1.10) are fitted with clustered standard error of date.

In table 1.9, I report the regression results associated with the predictive power of influencers' attention. The result shows a weak and positive relationship between the same day influencers' and WSB attentionand the coefficient is 0.0956 (t-stats 1.703). This indicates that when influencers mention a certain stock one more time, it can lead to a 0.0956 increase for the overall attention. The weak relationship is not persistent and I observe insignificant relationships between the WSB attention and influencers' 5-day lagged attention. This result suggests that influencer's attention is not a precursor of the WSB attention. Similarly, in table 1.10, although there is a more significant relationship between WSB sentiment and influencers' lagged sentiment, the coefficients tend to flip signs in a day or two. This may indicate disagreement, which can be a source of trading as mentioned in Pedersen (2021). On average, influencers' sentiment can not be a good predictor for the WSB sentiment.

Hypothesis 2: Reactions to Firms' Key Development Events

This hypothesis aims to test the responsiveness of influencers to firms' key development events. This can also be interpreted as whether influencers respond to firms' fundamental information. As examined before, Reddit WSB, as a community, does respond to most of the important key development events. To examine the relationship between WSB influencers' activities and firms' key development events, I conducted a similar regression as described in section 1.3.2. Specifically, the following model is examined:

$$\textit{Reddit Influencers' Activites}_{i,t} = b_0 + b_1 KeyDev_{i,t} + FE_{i,t} + e_{i,t} \tag{1.4}$$

The independent variable is WSB influencers' activities, and it can be measured by attention or sentiment. The attention is calculated as the daily number of stock mentions from WSB influencers, and sentiment is calculated as the average sentiment scores based

Table 1.9: This table reports results regarding the predictive power of WSB influencers. The dependent variable is Reddit WSB attention, measured by the daily total number of stock mentions in the "Daily Discussion" submissions. The independent variables include daily influencers' attention and its 5-day lags. The controls are stock return and volatility from $t-1$. Firm and date fixed effects are included. All regressions are fitted with a clustered standard error of date. Robust t-statistics are reported in parentheses. ***,**,* indicate the significance level of 1%, 5%, and 10%, respectively.

	(1)	(2)	(3)	(4)
VARIABLES	Reddit_Attn	Reddit_Attn	Reddit_Attn	Reddit_Attn
influencer_mention	0.102**	0.100*	0.0915	0.0956*
	(2.372)	(1.859)	(1.447)	(1.703)
influencer_mention (t-1)		-0.0172	-0.0335	-0.0448
		(-0.418)	(-0.492)	(-0.724)
influencer_mention (t-2)		0.00837	0.0212	0.0193
		(0.209)	(0.278)	(0.291)
influencer_mention (t-3)		0.00753	0.0271	0.0290
		(0.249)	(0.342)	(0.419)
influencer_mention (t-4)		-0.00337	0.0365	0.00412
		(-0.109)	(0.588)	(0.0790)
influencer_mention (t-5)		0.0256	0.0366	0.0208
		(0.799)	(0.592)	(0.477)
Observations	3,095,957	2,981,805	1,243,032	1,243,027
R-squared	0.005	0.005	0.005	0.145
Controls	No	No	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Date FE	No	No	No	Yes

Table 1.10: This table reports results regarding the predictive power of WSB influencers. The dependent variable is Reddit WSB Sentiment, measured by the sentiment score based on Laughran-McDonald Dictionary. The independent variables include daily influencers' sentiment and its 5-day lags. The controls are stock return and volatility from $t-1$. Firm and date fixed effects are included. All regression are fitted with a clustered standard error of date.Robust t-statistics are reported in parentheses. ***,**,* indicate the significance level of 1%, 5%, and 10%, respectively.

	(1)	(2)	(3)	(4)
VARIABLES	Reddit_Sent	Reddit_Sent	Reddit_Sent	Reddit_Sent
influencer_Sent (t)	-0.0389	-0.384	-0.689*	-0.150
	(-0.742)	(-1.173)	(-1.808)	(-0.338)
influencer_Sent (t-1)		0.0945	0.936*	0.706**
		(0.319)	(1.760)	(2.283)
influencer_Sent (t-2)		0.621**	0.675*	1.272***
		(1.987)	(1.911)	(3.192)
influencer_Sent (t-3)		-0.449**	0.968	0.965
		(-1.979)	(1.255)	(1.637)
influencer_Sent (t-4)		-0.0868	-2.192*	-2.196**
		(-0.265)	(-1.908)	(-2.298)
influencer_Sent (t-5)		-0.580	0.0642	0.521
		(-1.566)	(0.0712)	(0.521)
Observations	53,833	4,802	2,450	2,430
R-squared	0.024	0.105	0.118	0.351
Controls	No	No	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Date FE	No	No	No	Yes

on the Laughran-McDonald dictionary from WSB influencers. The independent variable *KeyDev* are the top 10 most frequent types of key developments on Reddit that happened for firm i on date t. For instances, Conferences, Earning Announcements, and Executive board changes. To account for firm- and time-specific variables, the firm and date fixed effects are added into the model, with a clustered standard error of date. If the coeffcient, b_1, is positive and significant, then Reddit influencers are shown to respond to the Key developments.

Table 1.11 reports the regression results. Each of the top 10 key developments is examined in the regression, however, the insignificant relationships are omitted in the table for better readability. On the "attention" front, influencers appear to respond to the majority types of key development events, similar to the general Redditors. However, the influencers' sentiment does not appear to respond to any type of key development events, except for the "Conferences". Particularly, influencers react positively when there is a "Conference", and the coefficient is 0.229 (t-stats 1.757). On the contrary, the WSB general public reacts negatively, with a coefficient -0.183 (t-stats -1.981). What's more, the general public responds negatively to "Earning Announcement", "Buyback Tranche Update", and "Corporate Guidance", while the influencers establish no significant relationship with them. This indicates that influencers' sentiment does not reflect the general public's sentiment, and on average, the influencers and non-influencers can hold different opinions, on the event day. The lack of responsiveness to the influencer's sentiment could indicate influencers are indeed hard-headed.

Hypothesis 3: Return Predictability

It is also unclear whether the influencers should establish a better or worse return predictability compared to the general Redditors. If the influencers are hard-headed rational investors, who influence many other naive investors to trade, then they should have some ability to predict stock returns, and their influences could be one of the explanations for the positive return predictability of WSB established in the previous section. However, it

Table 1.11: This table compares the coefficients from multiple similar regressions: $Reddit\ Activites_{i,t} = b_0 + b_1 KeyDev_{i,t} + FE_{i,t} + e_{i,t}$, the coefficients b_1 are documented below in the table. The independent variables can be the attention/sentiment from influencers or WSB. **Insigniciant coefficients are omitted for better readability**. Robust t-statistics are reported in parentheses. ***,**,* indicate the significance level of 1%, 5%, and 10%, respectively.

	(1)	(2)	(3)	(4)
	Attention		Sentiment	
VARIABLES	Influencer	WSB	Influencer	WSB
Conferences	**-0.00429*** **	**0.240** **	**0.229***	**-0.183** **
	(-2.960)	-2.021	(1.757)	(-1.981)
Company_Conference_Presentations	**-0.00243** **	-	-	-
	(-2.100)	-	-	-
Earning_Announcement	**0.00519***	**3.601*** **	-	**-0.267*** **
	(1.871)	-10.67	-	(-2.828)
Product_Related_Announcements	-	**0.994*** **	-	-
	-	-4.828	-	-
Earning_Calls_Time	**0.00389** **	-	-	-
	(2.233)	-	-	-
Earning_Release_Date	**0.00330***	-	-	-
	(1.752)	-	-	-
Executive_Board_Changes	-	**0.539*** **	-	-
	-	-3.586	-	-
Client_Announcements	**0.00315***	**1.150*** **	-	-
	(1.703)	(4.991)	-	-
Buyback_Tranche_Update	-	**2.684*** **	-	**-0.249** **
	-	-8.896	-	(-2.240)
Corporate_Guidance	**0.0107** **	**3.816*** **	-	**-0.387*** **
	(2.155)	-10.09	-	(-3.175)
Date FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes

appears that there is a dissociation between the influencers and general Redditors on day t, and influencers' activities can't be treated as a precursor/predictor for the general Redditors.

In the following table 1.12, I conduct similar regressions as in section 1.3.1. The predictive regressions of the return on a 10-day horizon are as the following:

$$Ret_{i,t+d} = b_0 + b_1 Reddit\ Influencer\ Measurements_{i,t} + b_3 Controls + FE_t + e_{i,t} \tag{1.5}$$

For stock ticker i on the day t, $Reddit\ Influencer\ Measurements_{i,t}$ could be (1) Reddit Influencers' Attention, measured by daily numbers of stock mentioned or (2) Reddit Influencers' Sentiment based on VADER, to stay consistent with the previous analysis.

The result shows that influencers establish a weaker predictive power for the majority of the 10-day returns. The influencers' attention can only predict return for day $t+5$, and the influencers' sentiment can only predict return for day $t+1$ and $t+8$. This result confirmed my previous results that influencers can neither control how the WSB general public thinks nor predict a positive return.

1.5 Conclusion

In this chapter, I present some stylized facts about the subreddit WallStreetBets and examined the information content of the subreddit. Additionally, I investigated whether the influencers on WSB are capable of predicting the aggregate opinion of the community and the stock returns.

To summarize, Reddit data is incredibly powerful and versatile. There are many different types of information other than texts, for instance, memes and other linked websites are popular on WSB. My study only includes textual information, and I created two types of measurements to gauge users' activities. Reddit Attention is constructed as the daily number of stock mentions, and Reddit sentiment is constructed as the average daily sentiment scores for the stocks.

Utilizing the above-mentioned measurements, I examined whether Reddit activities can

Table 1.12: This table compares the return predictability between WSB Redditors and influencers. Insignificant relationships are omitted for better readability. The regressions are $Ret_{i,t+d} = b_0 + b_1 Reddit\ Measurements_{i,t} + b_3 Controls + FE_t + e_{i,t}$. The controls are *ln(market cap)*, and *ln(bm)*, *ret (t-1)*, *ret (d2-d20)*, and *ret (d21-d120)*, which are natural logarithm of firm's market cap and book-to-market ratio, stock return on day $t-1$, cumulative stock return from day $t-20$ to day $t-1$, and cumulative stock return from day $t-120$ to day $t-21$, respectively. The regressions are fitted using panel regression with date fixed effect and clustered standard error of date. Robust t-statistics are reported in parentheses. ***,**,* indicate the significance level of 1%, 5%, and 10%, respectively. The more relevant coefficients are in bold.

	(1)	(2)	(3)	(4)
VARIABLES	Influencer_Attn	WSB_Attn	Influencer_Sent	WSB_Sent
Ret (t)	-	**-0.196***	-	**1.382****
	-	(-1.690)	-	(2.765)
Ret (t+1)	-	**-0.366***	**-0.0717***	**2.145****
	-	(-1.851)	(-1.664)	(2.444)
Ret (t+2)	-	**-0.375****	-	**3.497****
	-	(-1.968)	-	(3.625)
Ret (t+3)	-	-	-	**3.344****
	-	-	-	(2.626)
Ret (t+4)	-	**-0.301***	-	**1.813***
	-	(-1.925)	-	(1.956)
Ret (t+5)	**-0.0404****	-	-	**2.429****
	(-2.519)	-	-	(3.235)
Ret (t+6)	-	**-0.286****	-	**3.159****
	-	(-2.030)	-	-3.979
Ret (t+7)	-	**-0.595*****	-	**3.342****
	-	(-3.206)	-	(4.106)
Ret (t+8)	-	-	**0.0661***	**2.628****
	-	-	(1.893)	(3.048)
Ret (t+9)	-	-	-	-
	-	-	-	-
Ret (t+10)	-	**-0.434***	-	**2.608****
	-	(-1.935)	-	(2.706)

predict stock returns over a 10-day period of time, and concluded that the community as a whole can help predict the return. However, the predictive power of influencers is much weaker or even non-existed. Similarly, WSB reacts to most of firms' key development events, and influencers established a much weaker relationship. A further investigation between Reddit influencers and the general Redditors on WSB shows that influencers' activities can not be treated as a precursor or predictor for the entire community. This result indicates that the mechanism proposed in Pedersen (2021) may only apply to certain trading episodes, for example, the GameStop Short Squeeze event.

This chapter adds to the literature by providing more characteristics of Reddit WSB, and more importantly, by combining social networks with the financial market, I was able to show that Reddit influencers can not sway the entire WSB community. Instead of one-leads-other-follow, WSB aggregated activities reflect the wisdom of the crowd. Reddit, from an information perspective, does contain useful information, and it has the potential to be the "technology improvement" for retail traders.

Chapter 2

SOCIAL MEDIA, RETAIL TRADERS, AND VOLATILITY

2.1 Introduction

Over the past two decades, retail investors have changed tremendously. Individual tradings that used to be reserved only for some people are now open to everyone. Finance Literacy 101 classes are one click away. The Covid-19 pandemic and the hybrid working environment have allowed people more free time to trade. With all these changes, retail investors become a popular topic in the empirical finance literature. Retail investors, in this paper, are individuals who own individual brokerage accounts, come up with their own trading strategies, and usually place their trading orders on commission-free trading platforms. Many studies have shown that, historically, retail traders are disadvantaged when it comes to active trading. Barber and Odean (2000) documented that individual investors who hold common stocks directly pay a tremendous performance penalty for active trading, and overconfidence contributes to negative returns. There are also various studies focused on retail traders and the Internet. In Barber and Odean (2001), the authors mentioned that the internet changed how information was delivered to investors and how retail investors could act on that information. It has lowered the fixed and marginal cost of producing financial services and created challenges for the established providers of such services. Recently, many studies examined how various sources of information on the Internet affect retail investors' trading behavior. For example, Bali et al. (2021) found that among stocks dominated by retail investors, high analyst coverage and intense social interactions attract retail net buying. Farrell et al. (2022) looked at a crowd-sourced content provider, SeekingAlpha, and concluded that recent technology-enabled innovations in how individuals share information help retail investors become better informed. Notably, for the past few years, the social media platforms, Reddit

and Twitter, have become significant resources for understanding retail trading. For example, Rakowski et al. (2021) found that Twitter influences stock trading, and Twitter activity is associated with positive abnormal returns. Hu et al. (2021) concluded that higher Reddit activity, upbeat tone, and connectedness predicted higher returns in 2020-2021.

With the growing attention, it comes various different empirical results around the role of retail investors in the stock market. The emergence of fintech has led to the rise of Robinhood (or RH), a low-cost trading platform that has become popular among a new generation of retail investors, resulting in a surge of research on its impact on the stock market (Eaton et al., 2022; Ozik et al., 2021; Welch, 2020). One of the key questions in the literature is the role and impact of retail traders, including Robinhood users, on financial markets. The literature has established that noise traders should, in theory, contribute to volatility (De Long et al., 1990; Campbell and Kyle, 1993; Llorente et al., 2002). However, in the Robinhood literature, Eaton et al. (2022) shows that Robinhood traders increase idiosyncratic volatility while Ozik et al. (2021) finds that RH retail traders attenuate volatility in the stock market which, at first glance, seems contradictory. Our paper shows that the impact that RH traders have on risk, generally, depends on how we define risk. In particular, we show that RH traders increase idiosyncratic risk, which we define as the volatility of the residual estimated using various asset pricing models. Further, we show that RH traders actually decrease total volatility.

Our studyfootnoteJoint work with Anthony Sanford, HEC Montréal, Department of Finance. Assistant Professor. finds that both Eaton et al. (2022) and Ozik et al. (2021) are correct in their conclusions. Eaton et al. (2022) concludes that RH traders increase volatility by 0.238, whereas Ozik et al. (2021) finds that volatility decreases by about 0.001. Thus, there is apparent disagreement in the literature. However, we demonstrate in our study that the difference stems from the interpretation of the volatilities, not necessarily the results. Eaton et al. (2022) calculates volatility as the daily average of returns for 5-minute intervals, which is a highly noisy measure of volatility. In contrast, Ozik et al. (2021) measures volatility by simply calculating the daily standard deviation of returns. This seemingly

trivial difference is significant. As we demonstrate in our paper, Eaton et al. (2022) better reflects idiosyncratic volatility, given its short-term nature, while Ozik et al. (2021) more accurately reflects systematic risk. Therefore, both results are not necessarily contradictory but instead, complement each other in our understanding of the role or impact of RH traders on the market.

Our study also contributes to the already established literature that relates sentiments to volatility. For instance, Lee et al. (2002) demonstrates that sentiments have a significant impact on volatility. Other studies have also linked volatility jumps to sentiments (Sanford, 2022; John and Li, 2021; Borovkova and Mahakena, 2015; Seo and Kim, 2015). We propose to estimate the links between sentiments, risk, and RH trading. We estimate sentiments using data that was scraped from Reddit. Using this Reddit data, we then apply textual analysis to the posts in order to obtain sentiment indices for individual stocks on a given day. Using this data, we find that sentiments do indeed affect returns, idiosyncratic risk, and total stock risk.

Finally, our paper bridges the gap in the literature on noise trader characteristics. Specifically, we demonstrate that at very high frequencies, volatility is essentially a measure of idiosyncratic risk, which also contributes to the econometric literature related to risk. Our study demonstrates that one reason for the discrepancy in understanding noise traders as traders that attenuate or increase volatility lies in the estimation of volatility itself.

In summary, our study contributes to the literature on the impact of RH traders on financial markets by characterizing RH traders as traders that increase idiosyncratic risk but also decrease total risk. Moreover, we explore the determinants of the behavior of RH traders. We demonstrate that the differences in the results of previous studies on the impact of RH traders on volatility can be reconciled by accounting for the estimation of volatility itself. Finally, our study contributes to the broader literature on sentiments, volatility, and returns.

2.2 Models

In this section, we provide some theoretical exercises to motivate (1) the differences between idiosyncratic volatility and total volatility, (2) the relationship between social media sentiment and stock volatility, and (3) an overview of the regression specifications used.

2.2.1 Idiosyncratic Volatility

This paper adopts the Fama–French three-factor model to extract the idiosyncratic volatility of individual stocks[1]. We start by defining idiosyncratic volatility using a Fama-French three-factor model as in Ang et al. (2009). The three-factor Fama-French model is as the following:

$$r_{i,t} = \alpha_i + \beta_i MKT_{i,t} + s_i SMB_{i,t} + h_i HML_{i,t} + \epsilon_{i,t},\ \epsilon_{i,t} \sim (0, \sigma^2_{idio,i}) \tag{2.1}$$

where r_i is the return for firm i, $MKT_{i,t}$ is the market factor, $SMB_{i,t}$ is the small-minus-big factor representing the firm size premium, and $HML_{i,t}$ is the high-minus-low factor representing the value premium. The idiosyncratic volatility for firm i is thus the standard deviation of the residuals from equation 2.1, which we define as $\sigma_{idio,i}$. The regression is rolled on a monthly basis. To ensure the robustness of our results, we also estimate idiosyncratic risk using not only the F-F three-factor model but also the F-F 4 factor model as well as the Capital Asset Pricing Model (CAPM). These measures of idiosyncratic risk are all derived in the same way and so we do not include their derivations for the sake of brevity.

2.2.2 A Simple Example

To help explain the inconsistencies in the literature and further motivate our idea, we present a simple theoretical example in this section. As a reminder, Eaton et al. (2022) calculates volatility for a firm as the standard deviation of stock prices over a 5-minute interval, and, Ozik et al. (2021) calculates volatility as being the standard deviation of daily prices of firms'

[1]Fama/French 3 Research Factors, Kenneth R. French Data Library (`https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`).

stocks. What we aim to show in this section is that, Eaton et al. (2022), by calculating the standard deviation of high-frequency prices (5-minute intervals) is actually calculating the idiosyncratic risk, rather than total volatility. We will illustrate this idea using a simple theoretical exercise. For this exercise, we start with a model of idiosyncratic risk, by first calculating idiosyncratic return using the capital asset pricing model (CAPM) as follows:

$$r_{i,t} = \alpha_i + \beta_i[r_{m,t} - r_{f,t}] + \epsilon_{i,t} \tag{2.2}$$

which relates an individual firm's return, $r_{i,t}$, to that of the market's excess return, $r_{m,t} - r_{f,t}$. Idiosyncratic risk can then be defined as the standard deviation of the residual from this regression:

$$\sigma_{i,t}^{idio} = \sqrt{\frac{\sum_{t=1}^{T}(\epsilon_{i,t} - \bar{\epsilon_i})^2}{T-1}} \tag{2.3}$$

which, again, is nothing more than the standard deviation of the residual from the capital asset pricing model. Volatility, on the other hand, is generally defined as the standard deviation of returns for an individual firm's stock (in the case of an individual firm):

$$\sigma_{i,t}^{total} = \sqrt{\frac{\sum_{t=1}^{T}(r_{i,t} - \bar{r_i})^2}{T-1}} \tag{2.4}$$

which we call total volatility in this case in order to distinguish this measure from our previously defined measure of idiosyncratic volatility. Equation 2.4 is the method used by both Eaton et al. (2022) and Ozik et al. (2021). The difference, however, comes from the different horizons used to calculate each volatility measure. More specifically, Ozik et al. (2021) calculate their volatility measure using daily returns while Eaton et al. (2022) calculate their volatility measure using returns at the 5-minute interval. Although this difference may seem trivial, it is not. The reason for this is best shown using a simple example. Let us assume that markets are completely efficient. In particular, if we consider the average 5-minute time interval, we should not expect to have new information, on average. We start by re-writing equation 2.2 as:

$$\epsilon_{i,t} = r_{i,t} - (\alpha_i + \beta_i[r_{m,t} - r_{f,t}]) \tag{2.5}$$

which can be simplified by making the trivial assumption that over very short intervals, the risk-free rate is equal to zero, such that:

$$\epsilon_{i,t} = r_{i,t} - (\alpha_i + \beta_i[r_{m,t}]) \tag{2.6}$$

assuming that t is in minutes and by letting the return for the market be defined as a percentage change of the prices, we have the following relationship over a five-minute interval:

$$\epsilon_{i,t} = r_{i,t} - \left(\alpha_i + \beta_i \left[\frac{P_t - P_{t-5}}{P_{t-5}}\right]\right) \tag{2.7}$$

which can be further simplified by assuming that our traders are fully rational and that we have a well-behaved market such that the intercept of the market model is equal to zero:

$$\epsilon_{i,t} = r_{i,t} - \left(\beta_i \left[\frac{P_t - P_{t-5}}{P_{t-5}}\right]\right) \tag{2.8}$$

again, assuming that the market is efficient and that investors price additional information related to their market into their pricing equation as soon as it is available, our best guess as to what the difference between the price of the market five minutes ago and now is, in all likelihood, very close to zero (see, for example, Chan et al. (1991) and Laurent and Shi (2020) who show that index 5-minute returns are about 0.03% and 0%, respectively), therefore, we have:

$$\epsilon_{i,t} = r_{i,t} \tag{2.9}$$

assuming no new market information, all that we are left with is that, at very short intervals, we would expect price movements of individual assets to simply reflect innovations in the information of the firm, which would result in shocks to the stocks' return. As such, at very short intervals, we would expect that a model like the CAPM and volatility estimated from the individuals stock would reflect, on average, idiosyncratic risk such that $\sigma_{i,t}^{idio} \cong \sigma_{i,t}^{total}$.

As such, when we estimate volatility using a standard deviation of returns where the time window of the returns approaches 0 ($\Delta t \to 0$), what we have shown implies that this measure actually approximates our measure of idiosyncratic risk, not total volatility. In other words, we have shown that when the timing window for the return calculation approaches

zero, this implies that $\sigma_{i,t}^{idio} \cong \sigma_{i,t}^{total}$. A similar idea can also be conveyed in a full model (see Appendix). You can find a simulated result on the ratio of idiosyncratic volatility and total volatility in the next few pages.

2.2.3 Sentiment and Idiosyncratic Volatility

What about sentiments? Is there a relationship between trade conducted on the RH platform and the sentiments found on the social media platform Reddit? Both the media and finance literature point to a possible connection between social media (e.g. Reddit) and RH trades (Long et al., 2023; Malz, 2021; Hu et al., 2021). So, what is the relationship between Reddit user sentiment, RH traders, and risk? Theoretically, if we assume that there is a one-to-one relationship between Reddit activity and RH trading, then Reddit activity could be used as a proxy for the intent of RH traders and their overall trading direction. As such, we motivate one of our models in this paper using a theoretical framework from Mendel and Shleifer (2012). We start by defining a fundamental value for a hypothetical asset as:

$$V = \mu + \sigma_1\nu_1 + \sigma_2\nu_2$$

where μ is the unconditional expectation, $\sigma_1\nu_1$ is a shock that is realized in period 1, and where ν_1 is normally distributed with mean zero and variance one, and $\sigma_2\nu_2$ is a shock to the fundamental value which is only realized in period 2. We have three types of agents that participate in this particular market: noise traders, Insider/informed traders, I, and outsider/uninformed sophisticated traders, O. For this paper, we are focusing on RH traders and so we will be interested in determining if, as others in the literature have claimed, we can characterize RH traders as noise traders. We assume that this noise trader can be biased in their beliefs about the fundamental value of the asset and that that bias comes from a "sentiment" shock, S. The natural question, and the possible link we seek to explore in this paper, is whether or not RH trades can be traced back to Reddit sentiment shocks. To answer this question, we will model sentiments using the Reddit social media platform. Sentiments can be measured using various linguistic analysis tools (Loughran and McDonald, 2020). For

this paper, we have opted to use the method outlined in Loughran and McDonald (2011).

Specifically, in the Mendel and Shleifer (2012) model, we start by deriving demand for assets in two periods. The first period is the period of interest for the purposes of our hypothesis. The Mendel and Shleifer (2012) paper sets a more general framework where in period 1, investors trade their securities, and in period 2, the asset pays off its fundamental value. Since we are interested in determining the relationship between sentiments and the act of trading, we focus simply on the first period. Period-1 demand is derived from the utility maximization problem. In other words, agent i begins with a wealth W_i and will choose demand such that they maximize:

$$E_i\left[-e^{-\gamma(D_iV+(W_i-D_ip)r)}\right]$$

where E_i is the agent's expectation at time t, γ is a risk aversion parameter, D_i is agent i's demand for the asset, V is the fundamental value of the asset, p is the price of the asset, and r is the return on the asset. Using the result from the optimization problem, we can derive the agent's demand curve as being simply the first-order condition of the optimization problem as such:

$$D_i = \frac{E_i[V] - pr}{\gamma\sigma^2(V)}$$

where $E[\cdot]$ denotes the expectation with respect to agent i's information set and $\sigma^2(V)$ represents the variance of the fundamental value of the asset, V, conditional on agent i's information set. For the noise traders, the demand function then becomes:

$$D_{Noise} = \frac{\mu + S - pr}{\gamma\sigma^2_{Noise}}$$

Using the noise traders' demand function, we can then derive certain dynamics, one of which is the relationship between noise traders and sentiments:

$$\frac{\partial\sigma^2_{Noise}}{\partial S} = \frac{1}{\gamma D_{Noise}} > 0 \text{ iff } D_{Noise}, \gamma \geq 0$$

which, assuming positive demand and positive RA parameters, indicates a positive relationship. As such, this model posits that, if RH traders are indeed noise traders, one of their

characteristics should be that their demand for investments, and their trading behavior, should be positively associated with sentiment indices. If there is no relationship between trading in certain assets and sentiments, then we can, at least according to this model, assume that RH trader demand is fundamentally driven and that these traders are not, in fact, noise traders.

2.2.4 Regression Specifications

Our baseline regression is a panel regression with various measures of volatility as the dependent variable and RH users' holdings as the independent variable. Again, the goal of these baseline regressions is to establish the relationship between the changes in the RH holdings and their impact on various volatility measures. We, therefore, define the panel regression as:

$$\sigma_{i,t} = \beta_0 + \beta_1 RHH_{i,t} + D_t + D_i + \epsilon_{i,t} \tag{2.10}$$

where $\sigma_{i,t}$ is kept as a generic volatility variable, not total volatility, such that it can represent any volatility measure discussed earlier (the various idiosyncratic volatility measures and the total volatility), $RHH_{i,.t}$ is the RH user holdings for firm i at time t, D_t are the time fixed-effects, and D_i are the firm fixed-effects. After establishing the effects of RH users' holdings on volatility, we are interested in determining what drives these users' holdings. In particular, we are interested in whether RH users trade according to fundamental information (proxied using various information shock measures) or if they trade as if they were noise traders (e.g. trading on non-fundamental sources of information for unknown reasons). The current literature seems to point to the fact that RH traders do indeed hold 1) diversified portfolios, and 2) investments that have intrinsic value according to fundamental valuations (Welch, 2022; Barber et al., 2021). As such, we define our regression model for this research question as:

$$RHH_{i,t} = \beta_0 + \beta_1 \text{Info}_{i,t} + D_i + \epsilon_{i,t} \tag{2.11}$$

where $RHH_{i,.t}$ is the RH user holdings for firm i at time t, $\text{Info}_{i,t}$ are various measures of information shocks for firm i at time t, and D_i are the firm fixed-effects. The next regression specification is related to sentiments on social media (e.g. Reddit). In particular, we are interested in understanding the relationship (Mendel and Shleifer (2012)) between Reddit sentiments and idiosyncratic volatility, total volatility, and stock returns of individual firms. We test this, again, using panel regressions with various controls and fixed effects as follows:

$$\sigma_{i,t}^{idio} = \beta_0 + \beta_1 S_{i,t} + \beta_2 RHH_{i,t} + \sum_{z=1}^{N} \beta_{z,i,t} \text{Controls}_{z,i,t} + D_t + D_i + \epsilon_{i,t} \tag{2.12}$$

$$\sigma_{i,t}^{total} = \beta_0 + \beta_1 S_{i,t} + \beta_2 RHH_{i,t} + \sum_{z=1}^{N} \beta_{z,i,t} \text{Controls}_{z,i,t} + D_t + D_i + \epsilon_{i,t} \tag{2.13}$$

$$r_{i,t} = \beta_0 + \beta_1 S_{i,t} + \beta_2 RHH_{i,t} + \sum_{z=1}^{N} \beta_{z,i,t} \text{Controls}_{z,i,t} + D_t + D_i + \epsilon_{i,t} \tag{2.14}$$

where $\sigma_{i,t}^{idio}$ is the idiosyncratic volatility for firm i at time t, $\sigma_{i,t}^{total}$ is the total volatility for firm i at time t, $r_{i,t}$ is the return for firm i at time t, $S_{i,t}$ is the sentiment index obtained from social media (e.g. Reddit) for firm i at time t, $RHH_{i,.t}$ is the RH user holdings for firm i at time t, D_t are the time fixed-effects, and D_i are the firm fixed-effects. In some of the model specifications for the above regressions, we also include lagged sentiments in order to determine if there is a time-series effect of the sentiment variable on our various dependent variables.

2.3 Data and Summary Statistics

2.3.1 Robinhood Data

The RH data used in this paper is retrieved from the website `Robintrack.com`, which downloaded the hourly price and users' holding for every stock on RH from the RH application programming interface (API). The data is hourly ticker-level from May 2nd, 2018 to August 13th, 2020 for a total of about 12 million ticker-hour observations. The data stops on August 13th, 2020 because RH terminated all API connections thereby making it impossible

to obtain RH data after that point[2]. The data consists of three metrics: date/time, users' holdings, and ticker. Users' holding indicates the amount of RH accounts holding a particular stock on some day. On a typical day, for each ticker, the data will include one users' holding per hour and 24 entries per day. For example, if there are 500 RH accounts holding Amazon.com, Inc today at 6 am, the users' holding for $AMZN at 6 am today will be 500.

In this chapter, only the end-of-the-day users' holding is considered. The last entry, typically around 11 pm, is selected for each day into our dataset. In the end, Robinhood data has 5,891,775 ticker-day observations, with 8,595 tickers, from May 2nd, 2018, to August 13th, 2020. Interestingly, there is a sharp increase in RH users' holding around January 2020, the onset of the Covid-19 pandemic. This increase is possibly related to the removal of commission fees, information on social media, and more free time to retail traders due to the pandemic (Eaton et al. (2022), Hu et al. (2021), and Ozik et al. (2021)). The RH increase corresponds to Reddit's growing activities in 2020, see figure 2.1.

2.3.2 Reddit WSB Data

The Reddit WSB data used in this chapter comes from the Reddit application programming interface (API). We restricted the data to the subreddit *r/WallStreetBets* only and scrapped the following information from the WSB subreddit: title, date/time, body of submissions, and comments, along with the subreddit users' identification information. Submissions are the original content that users post to the site, while comments are responses to that content. Submissions and posts are the same things in our analysis. Using this data, we can get a sense of several interesting facts related to social media attention with respect to stocks. For example, we are able to ascertain how popular certain stocks are at any given time. We do this by looking at how many posts are posted about a given stock and how often and at what frequency people are discussing certain companies on social media. Figure 2.2 shows a time-series graph of the number of posts each day on Reddit and the number of stock-related

[2]The popularity information is returned in a more coarse form on the Robinhood website in 2022. The Robinhood Investor Index, website: `https://robinhood.com/us/en/investor-index/`

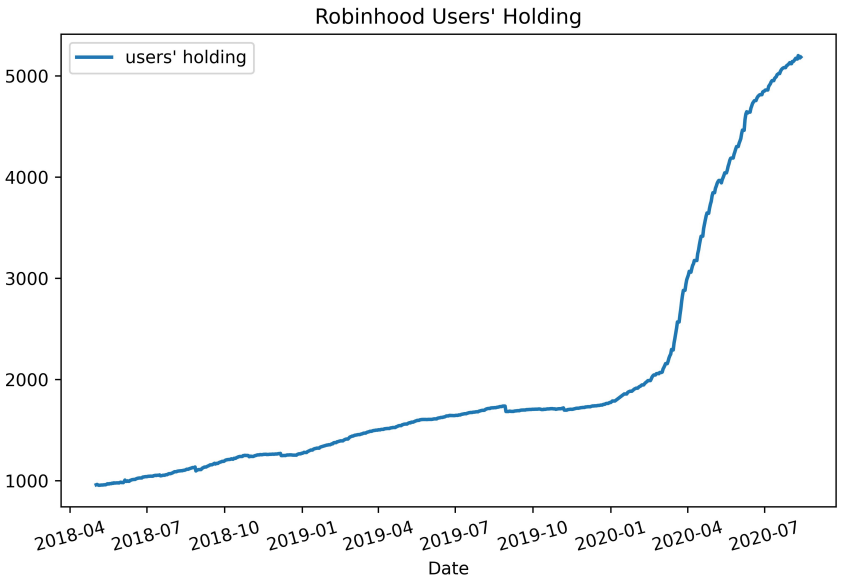


Figure 2.1: Robinhood Users' Holding Trend

posts on Reddit during our sample, respectively. Of particular interest from this graph is the amount of discussion taking place both on the WallStreetBets subreddit and the drastic increase of discussion related to stocks on the platform. Figure 2.2 shows an almost tenfold increase in the number of posts per day on the wallstreetbets subreddit during our sample alone. Figure 2.3, perhaps more importantly, shows that the variety of discussion about stocks has also drastically increased. At the beginning of the sample, WSB discussed less than 100 different companies on any given day. Towards the end of our sample, in late 2020, that number had increased to more than 500. An indication that the discussion about stocks on the WSB was clearly not simply limited to a narrow subset of stocks that are in the spotlight at any given but rather that the discussion was likely quite varied. The top 30 stocks discussed over our sample period can be found in the Appendix, to provide a sense of popular stocks discussed on Reddit WSB. The Reddit data used in this chapter is a subset of data used in the previous chapter, and interested readers can refer to the previous chapter for more information.

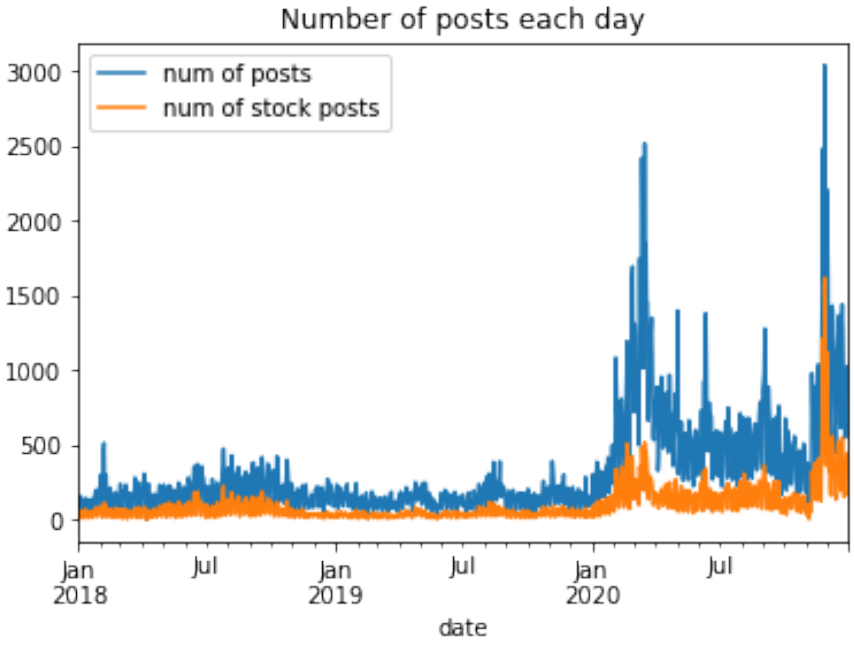


Figure 2.2: Reddit WSB's number of submissions each day. The blue line is the total number of submissions, and the orange line is the number of stock-related submissions.

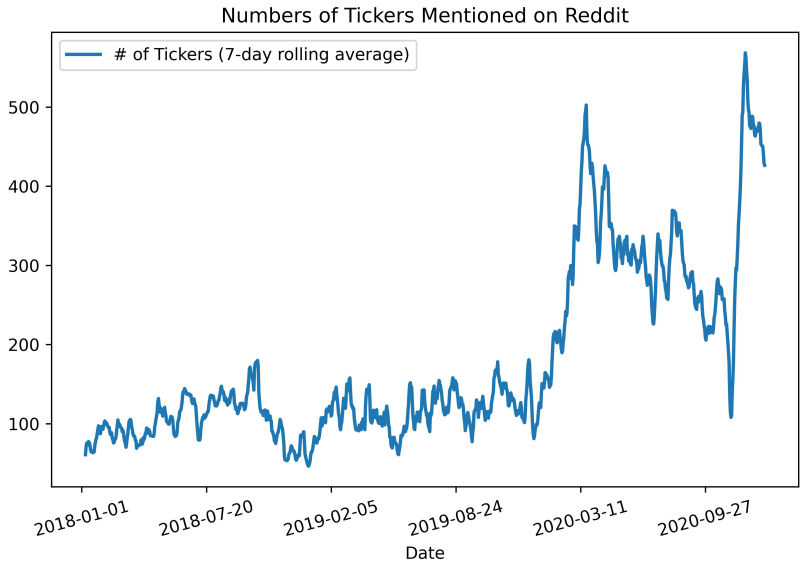


Figure 2.3: Reddit WSB's number of stock tickers mentioned, in 7-day rolling average.

From the Reddit data that we gathered, we were also able to obtain aggregate sentiments of Redditors using basic sentiment linguistic analysis techniques. In our case, we used the now ubiquitous method in the finance literature outlined in Loughran and McDonald (2011). The basic idea is to identify the sentiment words using the Loughran and McDonald (2011) dictionary and match the words with their corresponding sentiment scores. By subsetting Reddit discussions related to specific tickers, we can get a sense of the overall sentiment associated with a specific company's stock. The sentiment scores attached to each submission/comment are then aggregated for each day. For example, if the comments related to Apple were that the company was doing extremely well, that would be quoted as an overall positive sentiment. By aggregating discussions and sentiment scores on Reddit, we can get a sense, at any given time, of the overall market sentiment associated with a specific company. There are many alternatives when it comes to calculating the sentiment, for instance, using VADER (Valence Aware Dictionary and sEntiment Reasoner). More details can be found in the previous chapter.

2.3.3 Firm Data

The firm-level data used in this paper comes from three major sources: CRSP, Compustat, and Capital IQ, all of which are available via the Wharton Research Data Services (WRDS) portal. CRSP was used to obtain stock data for individual firms as well as market-level data. In particular, we obtained stock prices, returns, volume, and volatility. Value-weighted market return including distributions was also obtained from the CRSP database as a measure of market returns. Compustat was used to obtain firm-level fundamental data such as earnings and firm-level events. Firm-level events were obtained via Capital IQ which is available in the Compustat database. Capital IQ gives us dichotomous variables representing various key developments for individual firms. For example, we obtained earning announcements, which can have major impacts on stock prices, are key developments included in this database. This then allows us to determine if there is a relationship between how RH investors and key developments for a specific firm. The key development data can help us understand where

retail trading comes from and whether fundamental information plays a role in determining RH trading. Finally, we obtain Fama-French factors, used to estimate idiosyncratic volatility, directly from the Fama-French database, which is available directly via the WRDS platform.

2.3.4 Descriptive Statistics

Descriptive statistics for our main variables used in our analysis are presented in table 2.1. Of particular interest from the descriptive statistics is related to the distribution of the sentiment score variable which appears to skew towards the negative sentiment score. We can see this by looking at the mean, which is highly negative, but also the percentiles where the median is -1. The average return, at about 5%, is quite small, which makes intuitive sense since we are calculating daily returns for individual stocks. The distribution appears to be fairly symmetric, centering around 0 but with a slight negative skew. Users' holdings are the holdings of the RH traders for a given stock on a given day. We obtain this variable, as previously stated, by assuming that the holdings are those of a representative investor and use that as the overall volume of RH traders for that stock on that day. The users' holding variable has been log-transformed. Much like the return variable, nothing is particularly glaring here – the distribution appears to be quite symmetric around the mean. Finally, the FF3 idiosyncratic risk variable represents a firm's idiosyncratic risk on any given day. The variable is calculated as the standard deviation of the error term from a standard Fama-French 3-factor cross-sectional regression, as described in the previous section. To make interpretation easier in the regressions, we standardized the idiosyncratic variable. So the summary statistics represent the summary statistics for idiosyncratic risk standardized around the mean. The sentiment score, in this particular analysis, is the bottleneck in terms of sample size. One limiting factor to consider in our analysis is the sample size of the Reddit sentiment data (sample size is about 38 thousand daily observations). Since the sentiment is derived from the limited discussion of stocks on a given day on Reddit, the sample is quite a bit smaller than that of the rest of the data, including RH holdings, available. It is important to note, however, that this measure is the aggregate sentiment of all comments,

not the number of comments.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables	N	Mean	SD	Skewness	Kurtosis	P25	P50	P75
Sentiment Score	37,094	-0.600	2.743	-7.612	162.2	-1	-1	1
Return (in %)	8.434e+06	0.0497	4.097	30.20	4,881	-0.984	0	0.979
Users Holdings (log)	3.724e+06	5.215	2.103	0.143	3.075	3.784	5.209	6.607
FF3 Idiosyncratic Risk	1.017e+07	3.25e-10	0.999	2.321	13.01	-0.548	-0.225	0.297

Table 2.1: Descriptive statistics for the key variables used in the analysis of this paper. The sentiment score variable is the variable obtained from the sentiment linguistic analysis from Reddit chatter related to stocks. The return variable represents the return for the individual stocks analyzed in percent. Users' holding is the aggregated volume of trades for a given stock on a given day, on average, from Robinhood trades. The FF3 Idiosyncratic risk variable is the idiosyncratic risk (standardized cross-sectionally) for a firm obtained by taking the standard deviation of the residuals from the Fama-French factor model. The columns represent the sample size, mean, standard deviation, skewness, kurtosis, 25th percentile, median, and 75th percentile of the variables, respectively.

2.4 Results

The results section is divided into two sections: the simulated results and the empirical results. The simulated results are used to show, using simulated data, that the model derived 2.2.2 can be seen in a graph using simple data with known properties and basic definitions of risk. We then move onto our main analysis for the paper which uses empirical data.

2.4.1 Simulated results

Previously, we showed that the ratio of idiosyncratic risk to total risk will approach one when the time increment is small. Here, we show this using simulated data, under the CAPM model. Let t denote time and r_t denote the log return of the stock at time t. Similarly, let

m_t denote the log return of the overall market at time t. We assume that both r_t and m_t follow a normal distribution with mean and standard deviation as follows:

$$r_t \sim \mathcal{N}(\mu_r \Delta t, \sigma_r \sqrt{\Delta t})$$
$$m_t \sim \mathcal{N}(\mu_m \Delta t, \sigma_m \sqrt{\Delta t})$$

where μ_r and σ_r are the mean and standard deviation of the stock's returns, μ_m and σ_m are the mean and standard deviation of the market returns, and Δt is the time increment. We assume that the stock returns are a linear combination of the market returns and a firm-specific component:

$$r_t = \beta m_t + \epsilon_t$$

where β is the stock's beta (i.e., the sensitivity of its returns to changes in the market returns), and ϵ_t is the firm-specific component of the returns. We estimate the beta using a simple linear regression (CAPM-style model):

$$r_t = \alpha + \beta m_t + \epsilon_t$$

where α is the intercept for the model. The beta estimate is given by $\beta = \frac{\text{Cov}(r_t, m_t)}{\text{Var}(m_t)}$, where $\text{Cov}(r_t, m_t)$ is the covariance between the stock and market returns, and $\text{Var}(m_t)$ is the variance of the market returns. The idiosyncratic risk of the stock is defined as the variance of the firm-specific component of the returns, i.e., $\text{Var}(\epsilon_t)$. The systematic risk of the stock is defined as the variance of the market returns times the square of the beta, i.e., $\text{Var}(m_t)\beta^2$ (Sharpe, 1964). The total risk of the stock is defined as the variance of the stock returns, i.e., $\text{Var}(r_t)$.

We simulate the stock and market returns for a range of time increments Δt, and for each increment we calculate the idiosyncratic risk, systematic risk, and total risk of the stock. We then calculate the ratio of idiosyncratic risk to total risk for each time increment. The parameters used in the simulation are as follows: the mean and standard deviation of the market returns are $\mu_m = 0.05$ and $\sigma_m = 0.1$, respectively; the mean and standard deviation

of the stock returns are $\mu_r = 0.08$ and $\sigma_r = 0.2$, respectively; the correlation between the stock and market returns is 0.5; and the number of observations is 100. We simulate the time increment Δt over a range of values from 10^{-5} to 10^{3}. We then graph the relationship between time increment and the ratio of idiosyncratic risk to total risk to show that, as we stated previously, the ratio of idiosyncratic risk approaches one as the time increment used in the calculation of risk approaches zero:

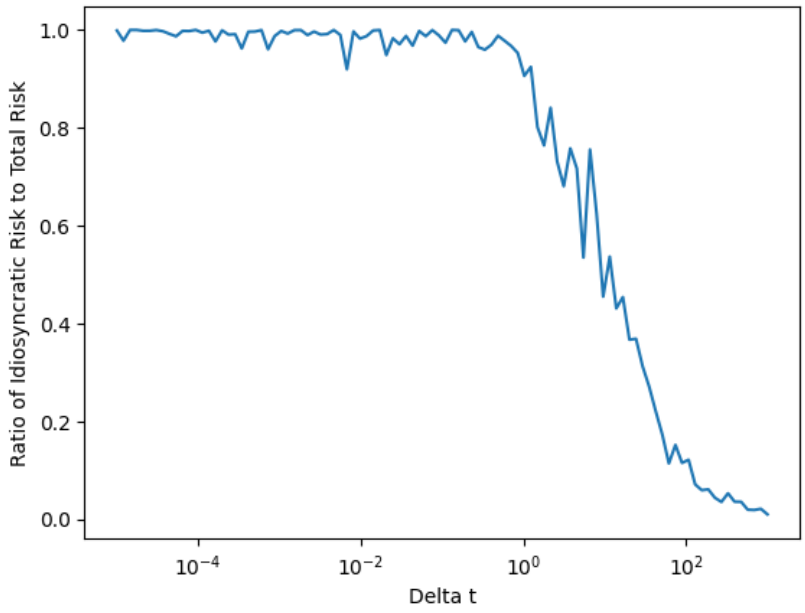


Figure 2.4: This figure shows the relationship between delta t and the ratio between idiosyncratic risk and total risk.

2.4.2 Empirical Results

Our first set of results can be seen in table 2.2. These are our baseline regressions which aim to ascertain the directional relationship between RH holdings and both idiosyncratic risk and systematic risk. This question is at the core of the disagreement between the findings of Ozik et al. (2021) and Eaton et al. (2022). On the one hand, Ozik et al. (2021)

finds a negative relationship between volatility and RH users holdings while Eaton et al. (2022) finds a positive relationship. Our first goal is to show that both findings are possible, depending on how we estimate and/or define volatility. Our first set of models has various volatility measures as dependent variables, the users' holdings from RH as the independent variable, and various other controls such as time and firm fixed effects. We find that there is a statistically significant and, perhaps more importantly, positive relationship between idiosyncratic volatility (models 1 through 3). We also find that there is a statistically negative relationship between RH users' holdings and total volatility. These results can be interpreted as RH users contributing to idiosyncratic volatility, but decreasing overall volatility for a particular firm on a given day. As such, this set of results shows that, indeed, both sets of results found in Ozik et al. (2021) and Eaton et al. (2022) are possible and that it simply boils down to our definition of risk to determine what the overall effect that RH traders might have on risk.

Table 2.3 shows the relationship between the RH users' holdings and various key corporate events obtained from the Capital IQ key developments database. The definitions of key development events can be found in the Appendix. The general idea for these regressions is to show that the RH users' holdings do, indeed, coincide with at least some firm-specific fundamentals. The events in capital IQ can be thought of as incidences where firms' fundamental information was either exchanged and/or changed. As such, we would expect investors to change their holdings in response to these events, which they indeed seem to be doing. In particular, RH users seem to change their holdings quite significantly in response to earnings announcements, earnings calls, earnings release dates, executive board changes, and buyback tranche updates. The signs on these coefficients also make intuitive sense. For example, we should expect a positive response to positive earnings announcements, which is what we find. We should also expect a negative response to executive board changes, generally associated with a signal of future uncertainty (Savor, 2012; Epstein and Schneider, 2008; Lin et al., 2003; Denis and Sarin, 1999), which is also what we find. In general, good news is associated with positive coefficients, and bad news or uncertain news is associated

	(1)	(2)	(3)	(4)
VARIABLES	Idio Vol FF3	Idio Vol FF4	Idio Vol CAPM	Total Vol
RH Users Holdings	0.000484***	0.000490***	0.000433***	-0.0000129***
	(8.204)	(7.736)	(8.355)	(-16.93)
Observations	2,271,690	2,270,853	2,273,476	2,378,565
R-squared	0.289	0.279	0.306	0.493
Time FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes

t-statistics in parentheses and errors are clustered at the date level.

*** p<0.01, ** p<0.05, * p<0.1

Table 2.2: Baseline regressions include RH Users Holdings as the dependent variable and various measures of idiosyncratic and systematic risk as independent variables. More specifically, the measures of risk included are, respectively: 1) idiosyncratic volatility derived from the standard deviation of the residual from the Fama-French 3-factor model, 2) idiosyncratic volatility derived from the standard deviation of the residual from the Fama-French 3-factor model and 3) idiosyncratic volatility derived from the standard deviation of the residual from the capital asset pricing model. The measure of systematic volatility is simply the volatility (measured as the standard deviation of returns) for an individual security. All models include firm and date fixed effects. Errors are clustered by date.

with negative coefficients. Moreover, the size of the coefficients is significant. For example, on days when earnings are announced, we would expect to see, on average, an increase in RH users' holdings of about 333, which is almost two standard deviations away from the mean. Executive board changes, on the other hand, cause a decrease in RH users' holdings, in all likelihood because of the increased uncertainty associated with new management (or the sign associated with changing the board), of about 177 which is a change of just about one standard deviation below the average.

Table 2.4 shows the relationship between idiosyncratic risk (calculated using the F-F3 specification) and various other dependent variables including sentiments, lagged sentiments, RH users' holdings, RH top 100 stock, and Reddit top 10 stocks. Sentiments represent the sentiment associated with certain stocks and/or companies on Reddit WSB at any given point in time. On any given day, for a firm, we aggregate the posts and topics that mention the company, convert their sentiments into scores using the methodology outlined in Loughran and McDonald (2011), and use that as our measure of sentiment with respect to a specific firm. This is, in essence, our empirical test for the notion put forth by Mendel and Shleifer (2012) that sentiments have an impact on noise. Note that here, our sentiment index is a dichotomous variable representing +1 if the sentiment is overall positive and -1 if the sentiment is overall negative. We chose this lexicon specifically because of its simplicity. Modern methods often introduce some subjectivity about the relative strength of sentiments. In our analysis, we were interested in determining if, on average and in general, sentiments could be associated with volatility. As such, we deliberately opted for a rather simplistic sentiment construction.

What we find is that, on average, sentiments do indeed increase idiosyncratic volatility. Our coefficient for the sentiment variable in the first model specification is highly statistically significant and is also quite sizable. We also find that our sentiment variable is robust to various other model specifications. We find that, on average, sentiments and idiosyncratic risk are positively related and that an increase in sentiment causes a one-point increase in overall sentiment.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	RH UH	RH UH	RH UH	RH UH	RH UH	RH UH	RH UH	RH UH	RH UH	RH UH
Conferences	-80.75									
	(-0.861)									
Company Conference		-137.9								
		(-1.545)								
Earning Ann			332.7**							
			(2.549)							
Product Related Ann				44.51						
				(0.447)						
Earning Calls					235.3*					
					(1.786)					
Earning Release Date						262.0*				
						(1.826)				
Executive Board Changes							-176.9*			
							(-1.660)			
Client Ann								78.76		
								(0.496)		
Buyback Tranche Update									429.1*	
									(1.812)	
Corporate Guidance										-3.493
										(-0.0239)
Observations	5,891,773	5,891,773	5,891,773	5,891,773	5,891,773	5,891,773	5,891,773	5,891,773	5,891,773	5,891,773
R-squared	0.680	0.680	0.680	0.680	0.680	0.680	0.680	0.680	0.680	0.680
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

t-statistics in parentheses and errors are clustered at the date level.

*** p<0.01, ** p<0.05, * p<0.1

Table 2.3: Panel regressions that include various key development indicator variables obtained from the Capital IQ database as the independent variables and the dependent variables are the RH users holdings. All models include firm fixed effects. Errors are clustered by date.

For robustness verification in our analysis, we also included two popularity measures: RH 100 and Reddit 10. Both of these are binary variables that represent popularity. In the case of the RH 100 variable, it equals 1 if the stock was in the Robinhood 100 most popular stocks and 0 otherwise. This is meant to see if the most popular or most highly traded stocks on RH behave differently in our analysis. This then ensures that our analysis is not driven simply by fringe stocks that are rarely traded on the RH platform. Moreover, we also include a similar index associated with Reddit, the Reddit 10 variable. This variable was constructed as one if the stock was one of the 10 most popular stocks on Reddit on any given day. The Reddit 10 measurement aims to supplement the RH 100 measurement and accounts for any stocks that are the new hits on social media but have yet made into the RH 100 Hall of Fame. Our results are robust to the inclusion of these variables. In particular, the coefficients on the sentiment index and the RH user holdings do not seem to be affected by subsetting our data to either of these controls.

Table 2.5 shows more or less the same regression as in Table 2.4 with the exception that here, our dependent variable is total volatility where volatility is defined simply as the standard deviation of returns. Again, our independent variables include: sentiments, lagged sentiments, RH users' holdings, RH top 100 stock, and Reddit top 10 stocks. Importantly, in these models, we note that the signs for the coefficients have switched to negative. The implications of this are that now, instead of increasing volatility, Reddit sentiments and RH users' holdings attenuate total volatility. This implies that, more specifically, an increase in the sentiment index leads to a decrease in a stock's total volatility. In general, this is an indication that positive sentiments reflect overall positive information in the market which decreases a stock's overall uncertainty. This is a particularly interesting result when we put this into perspective: table 2.4 shows that idiosyncratic risk increases with sentiments while results in 2.5 point to the fact that a firm's total risk decreases as a result of increase sentiments. By our definition of total risk, this would seem to indicate that a firm's exposure to systematic risk is "lessened" by overall sentiment. So sentiments increase firm-specific "noise" which, in turn, decreases the overall importance of the systematic component in the

VARIABLES	(1) Idio Vol	(2) Idio Vol	(3) Idio Vol	(4) Idio Vol	(5) Idio Vol	(6) Idio Vol
Sentiment	0.0103***	0.00828**	0.0139***	0.0114**	0.0139***	0.0139***
	(3.771)	(2.055)	(3.730)	(2.068)	(3.725)	(3.730)
Lagged Sentiment		0.00865**		0.00925*		
		(2.190)		(1.734)		
RH User Holdings			4.09e-07***	1.02e-06***	3.06e-07**	4.11e-07***
			(2.682)	(3.920)	(1.982)	(2.693)
RH 100					-0.399*	
					(-1.757)	
RH 100 X Sentiment					0.0364	
					(0.685)	
Reddit 10						0.472
						(1.539)
Reddit10 X Sentiment						0.0564
						(0.655)
Observations	22,381	6,561	14,928	4,393	14,928	14,928
R-squared	0.388	0.454	0.435	0.483	0.436	0.435
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes

t-statistics in parentheses and errors are clustered at the date level.

*** p<0.01, ** p<0.05, * p<0.1

Table 2.4: Panel regressions which consist of idiosyncratic risk as the dependent variable. The sentiment variable is a dichotomous (+1 or -1) variable representing the level of sentiment about a specific stock on a specific day on the Reddit social media platform. The lagged sentiment variable is a one-period lag on the sentiment variable. RH User holding is the volume of RH holdings for a given stock on a given day. All models include firm and date fixed effects. Errors are clustered by date.

total risk model previously described.

Table 2.6 shows, again, the same regressions as tables 2.4 and 2.5 with the exception that now we have the return for the firm as the dependent variable. Here we find both statistical and substantive results. In particular, we note that an increase of one in the sentiment score leads to an increase of about 0.065% in daily returns, a quite substantive result. Moreover, these results are robust to the inclusion of lags of sentiment (which are not statistically significant), RH users' holdings, and our popularity index. In fact, in these return regressions, the only statistically significant result is that from the sentiment score variable. Even the RH user holdings do not have a significant effect on stock returns, which is interesting in and of itself. In particular, the fact that the users' holdings from RH have no significant impact on returns is an indication that RH users as a whole are likely too small of a group to substantively affect the overall stock market returns. Although the importance of RH seems to have increased substantively, in particular during the Covid pandemic, they are still not a large enough mass of investors to play a significant role in sock returns, generally, based on our data. These results are consistent with the literature that argues that, in general, volume and holdings do not explain stock returns (Lee and Rui, 2002; Gallant et al., 1992) but these results are consistent with the notion that sentiments, themselves, might incorporate valuable information about the stock market and therefore could explain stock returns (Joseph et al., 2011; Baker and Wurgler, 2006; Sanford, 2022).

2.5 Conclusion

Our study makes significant contributions to the literature on the impact of retail traders on financial markets. The emergence of Robinhood and other fintech platforms has led to a surge of research on the role and impact of retail traders on financial markets, particularly in the stock market. Our study adds to this literature by characterizing RH traders as traders that increase idiosyncratic risk but also decrease total risk.

One of the key findings of our study is that the apparent disagreement in the literature on the impact of RH traders on volatility is due to the interpretation of the volatilities used

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Vol	Vol	Vol	Vol	Vol	Vol
Sentiment	-0.00520**	-0.00658*	-0.0104***	-0.00768**	-0.0105***	-0.0104***
	(-2.425)	(-1.929)	(-4.285)	(-2.029)	(-4.319)	(-4.270)
Lagged Sentiment		-0.00508		-0.0104***		
		(-1.644)		(-3.129)		
RH User Holdings			-1.24e-06***	-1.60e-06***	-1.25e-06***	-1.24e-06***
			(-8.159)	(-6.783)	(-7.896)	(-8.152)
RH 100					-0.0540	
					(-0.331)	
RH 100 X Sentiment					0.0126	
					(0.281)	
Reddit 10						-0.0980
						(-0.211)
Reddit10 X Sentiment						-0.157
						(-0.492)
Observations	25,390	7,864	16,975	5,254	16,975	16,975
R-squared	0.723	0.747	0.770	0.793	0.770	0.770
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes

t-statistics in parentheses and errors are clustered at the date level.

*** p<0.01, ** p<0.05, * p<0.1

Table 2.5: Panel regressions which have systematic risk (volatility) as the dependent variable. The sentiment variable is a dichotomous (+1 or -1) variable representing the level of sentiment about a specific stock on a specific day on the Reddit social media platform. The lagged sentiment variable is a one-period lag on the sentiment variable. RH User holding is the volume of RH holdings for a given stock on a given day. All models include firm and date fixed effects. Errors are clustered by date.

VARIABLES	(1) Return	(2) Return	(3) Return	(4) Return	(5) Return	(6) Return
Sentiment	0.0647***	0.0914***	0.0742***	0.104***	0.0734***	0.0742***
	(3.690)	(4.591)	(3.458)	(4.366)	(3.431)	(3.457)
Lagged Sentiment		-0.0152		-0.00394		
		(-0.880)		(-0.183)		
RH User Holdings			-5.99e-07	1.57e-06	-6.72e-07	-6.02e-07
			(-0.640)	(1.157)	(-0.735)	(-0.643)
RH 100					-0.322	
					(-0.464)	
RH 100 X Sentiment					0.293	
					(0.700)	
Reddit 10						-1.195
						(-1.247)
Reddit10 X Sentiment						-0.346
						(-0.752)
Observations	36,523	10,702	24,307	7,095	24,307	24,307
R-squared	0.236	0.261	0.256	0.333	0.256	0.256
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes

t-statistics in parentheses and errors are clustered at the date level.

*** p<0.01, ** p<0.05, * p<0.1

Table 2.6: Panel regressions which have return as the dependent variable. The sentiment variable is a dichotomous (+1 or -1) variable representing the level of sentiment about a specific stock on a specific day on the Reddit social media platform. The lagged sentiment variable is a one-period lag on the sentiment variable. RH User holding is the volume of RH holdings for a given stock on a given day. All models include firm and date fixed effects. Errors are clustered by date.

to measure risk. Specifically, Eaton et al. (2022) calculates volatility as the daily average of returns for 5-minute intervals, which is a highly noisy measure of volatility. In contrast, Ozik et al. (2021) measures volatility by simply calculating the daily standard deviation of returns. Our study shows that Eaton et al. (2022) better reflects idiosyncratic volatility, given its short-term nature, while Ozik et al. (2021) more accurately reflects systematic risk. Therefore, both results are not necessarily contradictory but instead, complement each other in our understanding of the role or impact of RH traders on the market. Furthermore, our study bridges the gap in the literature on noise trader characteristics by demonstrating that at very high frequencies, volatility is essentially a measure of idiosyncratic risk. This finding contributes to the econometric literature related to risk and provides insights into the behavior of noise traders in financial markets.

Our study also finds that RH traders trade based on shocks to fundamentals rather than solely relying on Reddit's platform as a mean for determining their trades. This finding challenges the popular belief that RH traders are solely influenced by social media sentiment and trade mostly meme stocks. By estimating the links between sentiments, risk, and RH traders, we find that sentiments do indeed affect returns, idiosyncratic risk, and total stock risk. However, our findings indicate that RH traders trade by using shocks to fundamentals as signals for entry and exit, similar to other informed traders in the market.

Overall, our study provides a more nuanced understanding of the impact of RH traders on financial markets by exploring the determinants of their behavior and the measurement of risk. Our findings contribute to the broader literature on sentiments, volatility, and returns, and provide important implications for investors, policymakers, and market regulators. By understanding the behavior of retail traders and the factors that drive their trading decisions, market participants can better anticipate market movements and make more informed investment decisions.

Chapter 3

TWO TYPES, TWO TALES: ROBINHOOD AND TAQ RETAIL TRADERS

3.1 Introduction

In the previous chapter, I focused on the Robinhood retail traders. Although Robinhood traders are gaining attention from both the financial industry and academics, it should be clear that Robinhood traders may not be representatives of the entire population of retail traders. In fact, on the contrary, Robinhood traders establish some unique characteristics when it comes to their information-gathering process and trading behaviors compared to the general retail population. For instance, Barber et al. (2021) discovers that some unique features such as "Top 10 Movers" on the RH app contribute to RH investors' stock trading, while the general retail traders do not always see such features on their trading platforms. Before diving into the differences, I first define general retail traders as individuals who own individual brokerage accounts with some other more traditional trading platforms that existed before Robinhood and trading platforms like Robinhood (for example, Fidelity). They also come up with their trading strategies and place their trading orders by themselves. In essence, they are the retail traders that were not captured by RH trading data described in the previous chapters. The two types of retail traders are reportedly very different. In 2021, the average age of RH investors is 31 years old, about 50% of them are first-timer traders, and the average account size is \$4,000[1]. Robinhood retail traders are significantly younger and own less wealth compared to general retail traders. For instance, on Charles Schwab, users have a much larger average age at 49 years old, and their average account

[1]https://time.com/nextadvisor/investing/brokerage-reviews/robinhood-review/

size is much bigger than Robinhood traders, at $234,000[2]. There is also a NASDAQ survey stating, "A generational shift is underway in how younger retail investors use digital and social media to do their own research and help them make informed decisions", and "Gen Z investors mostly use online discussion boards to gather stock trading information." Based on the aforementioned differences, I examine the RH investors and general retail investors (proxied by NYSE Trade and Quote data) under the influences of social media (proxied by Reddit WSB).

To investigate the differences between the two types of retail traders, I use data from (1) Robinhood API, as described in the previous chapter, and (2) use Boehmer et al. (2021) (or BJZZ(2021) as the authors preferred) to identify the general retail trading data from NYSE TAQ database. It is unlikely that the TAQ data contain RH trading unless the buyers of RH trading flow proceed to trade exactly like retail traders. The details can be found in the data section. A series of theoretical papers guide me through my analysis, for example, DeGroot (1974) presented a model which describes how a group of people might reach an agreement on a common subjective probability distribution for the parameter by pooling their individual opinions. A more recent work, Pedersen (2021) introduces rational agents and financial markets into an otherwise standard DeGroot (1974) model, and it provides some predictions when it comes to the interactions between the two types of retail traders. In the context of my analysis, I consider RH traders as fanatics, who only listen to themselves instead of learning from everyone in the market, and the TAQ traders as rational investors, who learn and update their beliefs all the time. The model predicts that once the RH traders form their beliefs, they will stop listening to anyone else. This can be linked to RH traders trading on Meme stocks that were frequently mentioned on Reddit WSB no matter how the firms performed. What's more, the Reddit WSB flair "YOLO" often indicates the irrationality behind the trading, or "I just love the stock". On the contrary, the TAQ traders, who learn from everyone, will observe RH traders or Reddit WSB, and potentially

[2]https://www.businessofapps.com/data/robinhood-statistics/

bet on network spillovers, which means they are likely to ride the bubble. This predicts that TAQ net trading can be positively related to lagged RH trading, but not the concurrent RH net trading. Then, the model proceeds to predict that the rational traders will bet on the reversals at some point, selling out their positions. Eventually, the fundamentals are revealed and fanatics will then gradually learn. This predicts that when there is concentrated trading or discussion on Reddit WSB or RH, TAQ trading can potentially predict positive stock returns, while the predictability of RH trading should be much weaker. Or, when all retail traders achieve losses, TAQ traders' losses is relatively smaller in magnitude compared to RH traders.

To examine the above-mentioned predictions, I conduct the following analysis - (1) Can Reddit WSB daily mentions predict TAQ or RH trading? This goes back to the assumptions I made about RH traders being fanatics and TAQ traders being rational. The results show that RH users' holding establishes positive and significant relationships with Reddit WSB mentions and its lags, while TAQ net trading does not, (2) Use TAQ net trading to predict the change of RH's users' holding and examine whether TAQ traders will ride on the bubbles, the result gives a relatively significant and positive relationship between TAQ net trading on day t and RH change of users' holding on t-4, and (3) to look at the two types of retail traders more holistically, I examine how their tradings of popular stocks on either RH or Reddit WSB can predict stock returns and stock idiosyncratic volatility. I define the popular stocks as the top 10 stocks mentioned on Reddit WSB daily, or the top 100 stocks traded on RH each day. The results show that RH trading on popular stocks is related to lower stock returns and higher idiosyncratic volatility, and TAQ trading on popular stocks is linked to higher stock returns and lower idiosyncratic volatility.

In the following sections, I will summarize the relevant studies (section 3.2), introduce the data (section 3.3) and present my analysis and results (section 3.4 and 3.5).

3.2 Related Literature

My study contributes to several streams of literature. The primary contribution is to the body of literature examining the newly available TAQ retail trading flow, and in particular how their behaviors differ from the RH traders. Boehmer et al. (2021) provides an easy method to identify marketable retail purchases and sales using the NYSE TAQ database, and they discovered individual stocks with net buying by TAQ retail investors outperform stocks with negative imbalances by approximately 10bps over the following week, and they further suggest that TAQ retail marketable orders might contain firm-level information that is not yet incorporated into prices. Greenwood et al. (2022) uses the algorithm of BJZZ(2021) to estimate stock-day-level measures of retail-initiated buys and sells and find that the stimulus check disbursement days saw increases in retail trading, particularly retail-initiated buy trades. On the similarity and differences between TAQ and RH traders, Barber et al. (2021) documents that both of the trading activities follow similar trajectories, with a marked increase in the pandemic period. It also shows that RH users show excessively concentrated trading activities, compared with the general population of retail investors as captured by the TAQ data set.Eaton et al. (2022) contrast outages at RH with outages at traditional retail brokers. They find that herding by inexperienced investors can create inventory risks that harm liquidity in stocks with high retail interest, while other retail trading improves market quality. My study adds to this line of literature by examining the interaction between the two types of retail traders under the influence of social media (proxied by Reddit WSB).

My study is also related to the line of research on social networks and retail trading. I use measurements derived from Reddit WSB and examine how attention and sentiment from Reddit can help predict retail trading, stock return, and idiosyncratic risks. Cookson et al. (2023) finds evidence of selective exposure to confirmatory information among 400,000 users on the investor social network StockTwits. beliefs formed in echo chambers are associated with lower ex-post returns, more siloing of information, and more TAQ trading volume. Bushee et al. (2020) examines how the media influences retail trade and market returns during

the "quiet period" that follows a firm's IPO, and finds that more media coverage during this period is associated with more purchases by TAQ retail investors and that such purchases are attention-driven, rather than information-based. Bradley et al. (2021) examines the market consequences of due diligence (DD) reports on Reddit's Wallstreetbets (WSB) platform. Over 2018-2020, they find that DD recommendations are significant predictors of one-month ahead returns, earnings forecast revisions, and earnings surprises. In addition, user comments are incrementally useful for predicting returns, and small TAQ retail trade informativeness increases following DD reports. However, all of these benefits reverse in the first half of 2021. Hu et al. (2021) uses Reddit data from 2020 to 2021 and documents Robinhood 50 stocks are more affected by social media activity, with stronger links among TAQ retail order flow, shorting flows and future returns. The major difference between my work and Hu et al. (2021) is I focus on the contrast between TAQ and RH traders, and how they interact under the framework of Pedersen (2021).

The findings in this chapter can also help provide some evidence on retail traders' "echo chamber" effect. In particular, Cookson et al. (2023) find that self-described bullish users are five times more likely to follow a user with a bullish view of the same stock than are self-described bearish users, using data from social network StockTwits. Many theoretical papers also point out that retail traders are subject to limited attention and often choose to focus on a few important decisions (Kőszegi and Matějka (2020)). These can be linked to my findings of RH investors trade in a much smaller ticker universe compared to the general population of retail traders, and their tradings' close association with social media attention and sentiment.

3.3 Data and Summary Statistics

3.3.1 TAQ Retail Trading Data

NYSE Trade and Quote (TAQ) database contains intraday transactions data (trades and quotes) for all securities listed on the New York Stock Exchange (NYSE) and American

Stock Exchange (AMEX), as well as Nasdaq National Market System (NMS) and SmallCap issues. Boehmer et al. (2021) detailed an algorithm to identify the retail trading flow using TAQ data - (1) transactions with a retail seller tend to be reported to a FINRA Trade Reporting Facility (TRF) and typically have an exchange code "D" in the TAQ consolidated dataset, (2) retail order flow is given a slight price improvement relative to the National Best Bid or Offer (NBBO). Thus the price improvement amounts are often 0.01, 0.1, and 0.2 cents. The TAQ database has adopted the Boehmer et al. (2021) algorithm and offered a tool to download the retail order flow. Interested readers can find more details about the algorithm and the TAQ data tool in the Appendix. I use general retail investors and TAQ investors interchangeably for the rest of the chapter. To match the period of the RH, the TAQ retail order flow is also from January 1st, 2018, to November 26th, 2020, on a daily basis. For the stock ticker i on the day t, the primary measurement of general retail trading is retail trading imbalances (Boehmer et al. (2021),Hu et al. (2021), and Eaton et al. (2022)), which is

$$mroib_{i,t} = \frac{Retail\ Buy\ Volume_{i,t} - Retail\ Sell\ Volume_{i,t}}{Retail\ Buy\ Volume_{i,t} + Retail\ Sell\ Volume_{i,t}}$$

The retail trading imbalances indicate the net buying behavior of the retail investors for ticker i on the day t. When the $mroib_{i,t}$ is high for some stocks, it can be interpreted as high net buying interest from the general retail investors.

3.3.2 RH Retail Trading Data

The RH data used in this chapter is the same as in the previous chapters. The detailed description is in section 2.3.1. To summarize, the RH data is collected from `Robintrack.com`, which retrieves hourly RH users' holdings from the RH API. User's holding represents the number of RH users who holds a particular stock on a given day. I took the last data entry on each day into my data set. In my case, the RH data has 5,891,775 ticker-day observations, with 8,595 tickers, from May 2nd, 2018, to August 13th, 2020.

One caveat is that RH trading activities may end up in the TAQ trading data and the

two types of retail trading may be mechanically related. My understanding is that this happens if and only if, Citadel, who fulfills almost all the RH orders proceed to trade like retail investors so that it will be picked up by the Boehmer et al. (2021) algorithm. It is impossible to quantify the magnitude of the potential overlapping between TAQ and RH trading, however, my results suggest that the two establish different patterns. Therefore, I believe that TAQ and RH data only overlapped on a few occasions and the two data sets should provide a nice contrast between different types of retail traders.

The two data sets do not provide an apple-to-apple comparison as TAQ provides trading data and RH only provides users' holding information. The best measurements that can allow a comparison between the two are (1) TAQ $mroib_{i,t}$, the marketable retail order imbalances, and (2) the daily change of RH users' holding for each stock on a given day. The two measurements are not directly comparable in magnitude, however, the trends established by them should reflect some information on retail traders' net buying.

3.3.3 Other Data

The firm-level data used in this paper comes from two major sources: CRSP and Compustat, all of which are available via the Wharton Research Data Services (WRDS) portal. CRSP was used to obtain stock data for individual firms as well as market-level data. In particular, we obtained stock prices, returns, volume, and volatility. Value-weighted market return including distributions was also obtained from the CRSP database as a measure of market returns. Finally, we obtain Fama-French factors, used to estimate idiosyncratic volatility, directly from the Fama-French database, which is available directly via the WRDS platform.

3.4 Models

In this section, I will present evidence and examine the hypotheses established in the introduction. Particularly, I will provide an overview of the two types of retail trading, including the ticker universe comparison and their connections with Reddit WSB in subsection 3.4.1 and 3.4.2). I also examine the relationship between RH and TAQ trading. Lastly, I present

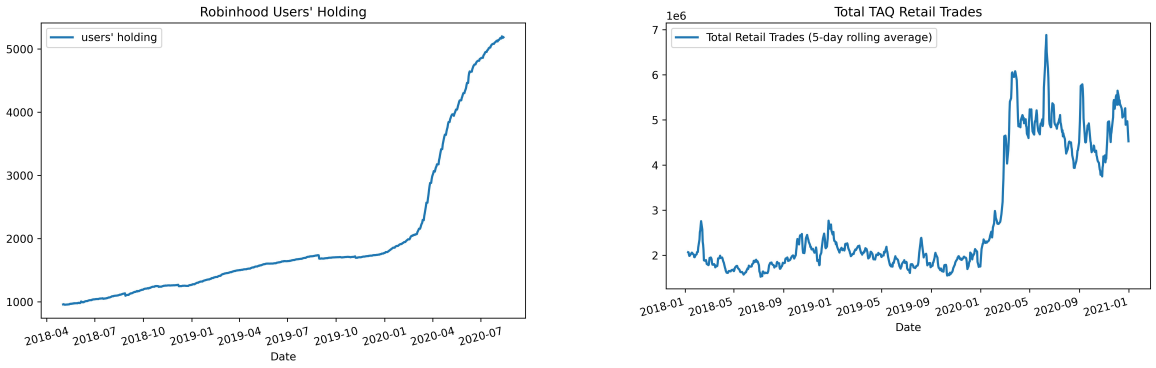
the return predictability of the two retail trading, and how their trading affects stocks' idiosyncratic volatility, especially for the popular stocks on Reddit WSB or RH.

3.4.1 Overview of the differences

I begin by outlining the similarities and differences between RH trading and retail trading reported in TAQ. As stated in the previous section, RH investors are generally characterized as younger, inexperienced, and possessing lower account balances than general retail traders. Nevertheless, these characteristics do not necessarily suggest that RH traders will respond differently to market information, particularly, information on Reddit WSB. Moreover, the smaller account sizes of RH traders may indicate that they have limited purchasing power in comparison to their counterparts in the TAQ dataset. These initial observations regarding RH and retail trading will serve as a basis for subsequent econometric modeling.

First of all, by plotting the RH daily average users' holding and TAQ total retail trades (number of retail orders for stocks on day t), I see the two measurements establish a very similar trend, with a sharp increase at the onset of the COVID-19 pandemic in early 2020. This corresponds to the widely reported surge in retail trader participation in the stock market during this period. RH is reported signed a record 3 million new users in the first four months of 2020, and in a 2021 Charles Schwab survey, it stated that fifteen percent of retail investors began investing in 2020.

Over 2018-2020, the RH ticker universe appears to be much smaller than TAQ's, and RH traders tend to trade cheaper stocks, potentially associated with their relatively small account size. RH stocks have an average price at \$37.25 (standard deviation \$84.77), while the average price of TAQ stocks is much larger, at \$50.65 (standard deviation \$1,578.49). For example, \$BRK, Berkshire Hathaway Inc Class A, priced at \$432,500, only appears in the TAQ retail stock pool. Specifically, there are around 1,500 tickers that are traded by TAQ retail traders only, and the average price of the 1,500 tickers is \$119.98, suggesting that more expensive stocks are less likely to appear in the RH ticker universe. The following figure shows some popular stocks in both ticker universes.



(a) RH **(b)** TAQ

Figure 3.1: Retail Traders' Growing Participation

(a) RH **(b)** TAQ

Figure 3.2: Popular Stock Tickers Traded

In the appendix, I present the top 30 commonly traded stock tickers in both the ticker universes. My analysis reveals that during the observed period, only 50% of the top 30 frequently traded TAQ tickers was also listed among the top 30 stocks for RH. Notably, popular TAQ stocks such as $AMC, $TQQQ, $SPY, $SNDL, and $BA did not make the top 30 list for RH investors. Given the premise that retail trading decisions are influenced by investors' information, wealth, and risk profiles, the results highlighted in this section provide suggestive evidence of heterogeneity between TAQ and RH investors.

3.4.2 Different Responses to Reddit r/WallStreetBets

Now that I have established that RH and TAQ traders are different in many ways, the next step is to understand how Reddit WSB influences overall retail trading. Reddit WSB is important in understanding the potential heterogeneity between TAQ and RH because not only it facilitates meme trading (Long et al. (2023)) but according to a NASDAQ survey, social discussion boards are among one of the major sources to gather information for the new generation of retail traders. The attention and sentiment expressed on WSB have led to significant changes in TAQ retail trading over 2020-2021 (Hu et al. (2021)). Nevertheless, the evidence is lacking when it comes to how TAQ and RH traders respond to the same Reddit WSB information.

The first set of results are between retail net buying and Reddit WSB attention measurement. I use the 5-day lags of Reddit WSB attention, accompanied by some control variables and date fixed effect, to predict the change of RH user's holding for stocks on day t. Then, the same regression is conducted with TAQ marketable imbalances for stocks on day t. The regressions are summarized below.

$$mroib_{i,t} = b_0 + \sum_{k=0}^{5} b_{k+1} WSB\ Attn_{i,t-k} + Controls_{i,t} + FE_i + e_{i,t} \tag{3.1}$$

$$Users'\ Holding(Log)_{i,t} = b_0 + \sum_{k=0}^{5} b_{k+1} WSB\ Attn_{i,t-k} + Controls_{i,t} + FE_i + e_{i,t} \tag{3.2}$$

Depending on the types of retail investors, the left-hand-side variable could be the TAQ retail trading order imbalances, $mroib_{i,t}$, or the natural logarithm of Robinhood users' holding, $Users\ Holdings(Log)_{i,t}$. $WSB\ Attn_{i,t}$ is the daily number of stock tickers mentioned, constructed based on the daily discussion submissions and comments.

The regression results in table 3.1 show that Reddit WSB momentum is significantly associated with the change of Robinhood retail trading and TAQ general retail order imbalances. However, the relationship between Robinhood retail trading and Reddit WSB momentum is much stronger and more persistent than with the TAQ general retail order imbalances. Specifically, Reddit WSB momentum is positively associated with the RH trading for all five-day lags, indicating that higher Reddit momentum relates to more intense buying behavior. The coefficient for the same day $Reddit\ Mom_{i,t}$ is 0.00312 with a 1% significance level when including 5-day lags, meaning that a 0.00312 increase of the Reddit Momentum for stock i on the day t is associated with 1 percent RH users' holding increase for the same day and stock. The other coefficients for $t-1, t-2, ..., t-5$ are all positive and significant, and the magnitude of the coefficients gradually decreases for a while and then increases. When looking at a same-day regression, the coefficient between Reddit Momentum and RH users' holding is 0.00856^{**}. On the flip side, the general population of retail investors' net buying behavior showed a negative and short-lived (1-day) relationship with Reddit momentum, indicative of contrarian trading. The coefficient of $Reddit\ Mom_{i,t-1}$ is negative, -0.000848 with 1% significance. This can be interpreted as when the Reddit Momentum for stock i on the day $t-1$ increases by 1, the TAQ retail order imbalances will decrease by -0.000848. TAQ traders will sell tickers with higher Reddit Momentum on average. This result suggests that Reddit momentum could influence both RH and TAQ retail investors' net buying behavior, but the effect is the opposite. The full regression tables are in the appendix table A.7.

To summarize, the result suggests that RH investors tend to be the net buyers of popular stocks on Reddit, measured by Reddit Momentum, while general retail traders tend to be net sellers. This result confirms one of the important RH traders' characteristics - trading attention-grabbing stocks, using Reddit-based attention measurements. Welch (2020),

Table 3.1: This table examines the relationship between retail trading and Reddit momentum. The dependent variable of the model (1) - (2) is the natural logarithm of daily Robinhood users' holding. The dependent variable of the model (3) - (4) is $mroib_{i,t}$, the marketable retail order imbalances. *WSB Attn* is the daily numbers of stock mentions. Firm characteristics controls (natural logarithm of the firm's market capitalization and natural logarithm of the book-to-market ratio of time t) are added as controls. All models are fitted using panel regression with date fixed effect and clustered standard error of date. Robust t-statistics are reported in parentheses. ***,**,* indicate the significance level of 1%,5%, and 10%, respectively.

	(1)	(2)	(3)	(4)
VARIABLES	*Users Holding*	*Users Holding*	*mroib*	*mroib*
WSB Attn (t)	0.00856***	0.00312***	-0.000213	0.000400
	(16.99)	(3.329)	(-1.073)	(1.258)
WSB Attn (t-1)		0.00321***		-0.000848**
		(2.985)		(-2.401)
WSB Attn (t-2)		0.00232*		-0.000112
		(1.933)		(-0.251)
WSB Attn (t-3)		0.00209*		0.000238
		(1.957)		(0.553)
WSB Attn (t-4)		0.00309***		4.71e-05
		(2.614)		(0.106)
WSB Attn (t-5)		0.00353***		-0.000556*
		(3.873)		(-1.780)
Date FE	Yes	Yes	Yes	Yes
Firm Controls	Yes	Yes	Yes	Yes
Observations	809,216	563,950	530,351	339,490
R-squared	0.003	0.003	0.002	0.002

Ozik et al. (2021), and Barber et al. (2021) also document that RH investors tend to trade attention-grabbing stocks using other attention measurements. Findings on the general retail traders are consistent with the notion that retail investors are contrarian traders (Eaton et al. (2022)), and they establish a different relationship with Reddit from RH traders. The result could explain the heterogeneity in the literature when it comes to whether retail investors are contrarian or not. When we subgroup the retail investors, some are seeking high-attention stocks, while others are avoiding them.

3.4.3 Connections between TAQ and RH Net Buying

As stated in the introduction, Pedersen (2021) implies that rational investors will observe fanatic investors trading on certain stocks and "ride on the bubble" at some point. In this section, I examine whether TAQ traders establish such behavior on average. It is possible that rational investors will not follow up with fanatics every step of the way, and they only act when they deem profitable (for example, when they think there is a bubble). If that is the case, I should not observe a significant and persistent relationship between RH and TAQ trading. The regressions are summarized below.

$$mroib_{i,t} = b_0 + \sum_{d=0}^{5} b_{1+d} Users' Holding(Log)_{i,t-d} + Date\ FE + Tic\ FE + e_{i,t} \tag{3.3}$$

Where *RH users' holding*$_{i,t}$ is the natural logarithm of one plus the daily RH users' holding of stock i on the day t, and $mroib_{i,t-d}$ is the general retail order imbalances of stock i on the day t. d represents the time lags added in the regression. The regression can be interpreted as using RH users' holdings on the day t and its 5-day lags to predict TAQ retail order imbalances on the day t. The association between the two types of retail investors could be seen through the coefficient b_{1+d}. If, over the course of one week, the net buying interests of TAQ and RH investors are somehow related, then I should observe a significant b_i. In table 3.2, I report the result of the equation 3.3. In column (1), I show that, on the same day, RH users' holding has no relationship with TAQ retail order imbalances. Even when I

include 3-day time lags, the association is still insignificant. When including more time lags of RH users' holding, there is one positive and significant coefficient (0.0262^{**}) between TAQ retail order imbalances on the day t and RH users' holding on the day $t-4$. This means that when RH users' holding for stock i increases by 1 percent on $t-4$, on average, TAQ retail order imbalances would increase by 0.0262 on the day t.

The fact that out of 5-day lags, there is only one positive and significant association between TAQ and RH net buying interests suggests that RH and TAQ traders do not invest in sync, and it isn't easy to use one to predict the other. It may suggest that TAQ observe and "ride the bubble", but the evidence is only suggestive. One can attempt to conduct the same analysis using "bubble" periods only, however, in this section, I am only interested to see that on average whether TAQ, the rational investors, would observe and act on RH traders, the fanatics. There is no clear evidence suggesting RH trading is associated with TAQ trading, and the disassociation can also add to the heterogeneity between RH and TAQ.

3.4.4 Informed or not, it's a question

My last question in this chapter goes back to the notion that RH traders are generally uninformed, and TAQ traders are better informed than RH traders. There is ample evidence showing that RH traders are subject to many behavioral biases, such as meme stock trading and trading attention-grabbing stocks. Their return after concentrated trading is often negative. On the contrary, Boehmer et al. (2021) documents positive returns for TAQ traders. As mentioned in the previous chapter, there are conflicting results when it comes to categorizing the market implications of retail trading. Furthermore, Pedersen (2021) suggests that fanatic investors, which are RH traders in my setting, only listen to themselves and trade on stocks they prefer. TAQ observes and can choose to follow RH traders when they see fit.

In this section, I examine how the retail trading around the popular stocks (on WSB and RH) is associated with stock return and idiosyncratic volatility. Many contemporary studies also focused on retail trading's return predictability to characterize the retail trading

Table 3.2: This table shows the relationship between RH retail trading and general retail trading order imbalances. The dependent variable of model (1)-(6) is $mroib_{i,t}$, the marketable retail order imbalances, and it is calculated by $mroib_{i,t} = \frac{Retail\ Buy\ Volume_{i,t} - Retail\ Sell\ Volume_{i,t}}{Retail\ Buy\ Volume_{i,t} + Retail\ Sell\ Volume_{i,t}}$. All models are fitted using panel regression with two-way fixed effects and clustered standard error of date. Robust t-statistics are reported in parentheses. ***,**,* indicate the significance level of 1%, 5%, and 10%, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	*mroib_t*	*mroib_t*	*mroib_t*	*mroib_t*	*mroib_t*	*mroib_t*
Users Holding (t)	0.000283	-0.00307	-0.00478	-0.00389	-0.00364	-0.00277
	(0.415)	(-0.526)	(-0.795)	(-0.644)	(-0.595)	(-0.448)
Users Holding (t-1)		0.00325	0.00276	0.00391	0.00189	0.000555
		(0.554)	(0.299)	(0.405)	(0.196)	(0.0556)
Users Holding (t-2)			0.00228	-0.0101	-0.00738	-0.00715
			(0.331)	(-0.967)	(-0.692)	(-0.652)
Users Holding (t-3)				0.0102	-0.00688	-0.0117
				(1.388)	(-0.582)	(-0.976)
Users Holding (t-4)					0.0162*	0.0262**
					(1.884)	(2.144)
Users Holding (t-5)						-0.00495
						(-0.640)
Observations	1,869,630	1,852,882	1,839,377	1,829,899	1,823,257	1,816,591
R-squared	0.016	0.016	0.017	0.017	0.017	0.017
Date FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes

behavior (Boehmer et al. (2021) and Barber et al. (2021)). What's more, idiosyncratic volatility can also be suggestive of the informativeness of retail trading, as stated in the literature (Aabo et al., 2017; Bekaert et al., 2012; Yang et al., 2020).

Therefore, I include stock return and idiosyncratic volatility in my analysis. To better characterize and contrast between RH and TAQ trading, I examine the relationship between (1) idiosyncratic volatility and RH retail trading, interact with indicators of popular stocks, (2) idiosyncratic volatility and TAQ general retail trading, interact with indicators of popular stock (3) return and interaction term of RH retail trading and popular stock indicators, and (4) return and TAQ general retail trading interacted with popular stock indicators. The popular stock indicators are created as dummy variables that equal 1 if stocks are among the popular stocks, and 0 otherwise. Particularly, I created - (1) Reddit top-10 list, which contains the top 10 most discussed stock tickers, measured by number of stock ticker mentions on a given day on WSB, and (2) RH top-100 list, which includes the top 100 stocks ranked by RH users' holdings. To provide a sense of the tickers that make the lists, see Table A.4, Table A.5, and Table A.6.

Idiosyncratic Volatility

This subsection connects stock idiosyncratic volatility $IVOL(t)$ with Reddit WSB attention and retail trading on the day t. The econometric specifications are as follows:

$$\begin{aligned} IVOL_{i,t} = b_0 &+ b_1 WSB\ Attn_{i,t} + b_2 RH\ Users\ Holding_{i,t} + b_3 Reddit10_{i,t} \\ &+ b_4 WSB\ Attn_{i,t} \times Reddit10_{i,t} + b_5 RH\ Users\ Holding_{i,t} \times Reddit10_{i,t} \\ &+ b_6 Controls + FE_{i,t} + e_{i,t} \end{aligned} \quad (3.4)$$

$$\begin{aligned} IVOL_{i,t} = b_0 &+ b_1 WSB\ Attn_{i,t} + b_2 RH\ Users\ Holding_{i,t} + b_3 RH100_{i,t} \\ &+ b_4 WSB\ Attn_{i,t} \times RH100_{i,t} + b_5 RH\ Users\ Holding_{i,t} \times RH100_{i,t} \\ &+ b_6 Controls + FE_{i,t} + e_{i,t} \end{aligned} \quad (3.5)$$

For the stock ticker i on the day t, $WSB\ Attn_{i,t}$ is measured by daily numbers of stock mentions (standardized cross-sectionally)[3], $RH100_{i,t}$ is the Robinhood top 100 list, $Reddit10$ is the Reddit top 10 list. $Controls$ are $ret_{i,t-1}$, cumulative return from day -20 to day -2, $ret_{i,d1-d20}$, and cumulative return from day -21 to day -120, $ret_{i,d21-120}$. Date and ticker fixed effects are added. For the TAQ retail traders, I simply replace $RH\ Users\ Holding_{i,t}$ with $mroib_{i,t}$, the TAQ marketable retail order imbalances. All regressions employ clustered standard error of date.

The idiosyncratic volatility regressions confirm the heterogeneity between RH and TAQ traders and show the opposite effect when it comes to retail trading around the top 100 popular stocks on RH. Particularly, RH trading on the RH100 stocks is positively and significantly associated with firms' idiosyncratic volatility, while TAQ trading on the same group of stocks gives negative and significant coefficients. For both types of retail traders, when considering WSB attention, retail trading, and stock idiosyncratic volatility together, WSB Attention and its interaction terms with retail trading show similar coefficients in terms of magnitude and significance.

I present a summary of the regressions in table 3.3. The full regression tables can be found in the appendix. In table 3.3 columns (1) and (3), I consider the two types of retail trading, interacted with Top 10 Reddit stock tickers. In column (1), the coefficients of *WSB Attn* is 0.00148, with 1% significance, and interaction terms of *Reddit10* and *WSB Attn* is 0.239 (1% significance). Similarly, in column (3), the coefficients of *WSB Attn* for TAQ traders is 0.00168, with 1% significance, and interaction terms of *Reddit10* and *WSB Attn* is 0.217 (1% significance). The coefficients suggest that for all stock tickers, higher WSB attention of the ticker relates to a higher firm's idiosyncratic volatility on the same day. What's more, when isolating the top 10 most mentioned stock tickers of the day, the effect is amplified, and the coefficients of the interaction terms are of much larger magnitude compared to other stock tickers. Notably, the coefficients of WSB-related measurements from the two regressions are

[3]Based on submissions titled "Daily Discussion" or "What Are Your Moves", and their comments.

Table 3.3: All models are fitted using panel regression with firm and date fixed effects and clustered standard error of date. Robust t-statistics are reported in parentheses. ***,**,* indicate the significance level of 1%, 5%, and 10%, respectively. Key results are highlighted.

	RH		TAQ	
	(1)	(2)	(3)	(4)
VARIABLES	*IVOL (t)*	*IVOL (t)*	*IVOL (t)*	*IVOL (t)*
Retail Trading	**0.00816****	0.00574	-0.000600	0.000594
	(1.977)	(1.390)	(-0.185)	(0.180)
Reddit10	0.154		-0.0311	
	(0.872)		(-0.750)	
Reddit10 × Retail Trading	-0.0248		0.0464	
	(-0.984)		(0.291)	
WSB Attn	**0.00148*****	0.00148***	**0.00168*****	0.00168***
	(3.450)	(3.470)	(4.130)	(4.146)
Reddit10 × WSB Attn	**0.239*****		**0.217*****	
	(3.157)		(3.656)	
RH100		**-1.410*****		**0.181*****
		(-6.384)		(9.787)
RH100 × Retail Trading		**0.253*****		**-0.107*****
		(7.331)		(-2.636)
RH100 × WSB Attn		-0.0323*		-0.0125
		(-1.742)		(-0.541)
Observations	476,819	476,819	627,114	627,114
R-squared	0.169	0.169	0.148	0.148
Date FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes

similar, suggesting WSB likely influences the entire population of retail traders. The RH retail trading in column (1) shows a positive and significant coefficient (0.00816 with t-stat 1.977), and this can imply RH retail trading is associated with higher firms' idiosyncratic volatility, however, this particular result is relatively weaker than other results as I did not observe the significance in other columns of the table. In summary, when looking at columns (1) and (3), I find WSB Attention and its interaction terms with top 10 Reddit stocks are positively associated with firms' idiosyncratic volatility, and this effect can be found for both types of retail traders. This also means WSB Attention can contain lots of noise, especially for the most discussed stock tickers, and the attention from Reddit adds to stock's risk.

In columns (2) and (4), I look at the RH 100 list, which contains the top 100 most held stock tickers, measured by the end-of-day Rh users' holding. Zooming in on the coefficients of *RH100*$\times$*Retail Trading*, I see the opposite effect for RH and TAQ traders. The RH retail tradings around the top 100 RH stocks establish a positive and significant relationship with firms' idiosyncratic volatility. Specifically, the coefficient is 0.253^{***} with a t-statistics at 7.331. On the contrary, the TAQ coefficient of the same interaction term gives a negative and significant coefficient, -0.107^{***} with a t-statistics at -2.636. In other words, on average, when trading the RH100 stocks, the RH investors' trading adds to the firms' idiosyncratic risk, while the TAQ trading does not. This also speaks to the informativeness of RH and TAQ and provides some evidence that RH is less informed than TAQ investors.

For robustness checks, to account for variations brought by different ways of calculating firms' idiosyncratic volatility, I also constructed $IVOL$ based on Fama-French 4-factor model and CAPM. The results can be found in the Appendix. The robustness checks show no major changes in my findings.

Return

This subsection connects daily stock returns with Reddit WSB Attention and retail trading on the day t. The econometric specifications are as follows:

$$Ret_{i,t} = b_0 + b_1 WSB\ Attn_{i,t} + b_2 RH\ Users\ Holding_{i,t} + b_3 Reddit10_{i,t} + b_4 WSB\ Attn_{i,t} \times Reddit10_{i,t} + b_5 RH\ Users\ Holding_{i,t} \times Reddit10_{i,t} + b_6 Controls + FE_{i,t} + e_{i,t} \quad (3.6)$$

$$Ret_{i,t} = b_0 + b_1 WSB\ Attn_{i,t} + b_2 RH\ Users\ Holding_{i,t} + b_3 RH100_{i,t} + b_4 WSB\ Attn_{i,t} \times RH100_{i,t} + b_5 RH\ Users\ Holding_{i,t} \times RH100_{i,t} + b_6 Controls + FE_{i,t} + e_{i,t} \quad (3.7)$$

For the stock ticker i on the day t, $WSB\ Attn_{i,t}$ is measured by daily numbers of stock mentions (standardized cross-sectionally), $RH100_{i,t}$ is the Robinhood top 100 list, $Reddit10$ is the Reddit top 10 list. $Controls$ are $ret_{i,t-1}$, cumulative return from day -20 to day -2, $ret_{i,d1-d20}$, and cumulative return from day -21 to day -120, $ret_{i,d21-120}$. Date and ticker fixed effects are added. For the TAQ retail traders, I simply replace $RH\ Users\ Holding_{i,t}$ with $mroib_{i,t}$, the TAQ marketable retail order imbalances. All regressions employ clustered standard error of date.

This set of results complements my previous findings. When using stock return as the dependent variable, column (1) and (3) show similar coefficients no matter which type of retail traders are considered. Particularly, although the *WSB Attn* on average establish positive relationships with stock return, the interaction terms between *WSB Attn* and *Reddit10* give negative coefficients of much larger magnitude. Specifically, for stocks that are not Reddit10, when they get one more mention on Reddit WSB, the stock return on day t is estimated to increase around 0.5-0.6 bps. However, if the stock getting more mentions belongs to the Reddit10 list, each mention will correspond to -0.372^{***} and -0.457^{***} change of stock return on day t, for RH and TAQ traders, respectively.

The results around RH100 stocks imply the same takeaway as before. *Retail Trading*$\times$*RH100* is -0.149^{**}, the t-statistics is -2.395, for RH traders. The coefficient of the

Table 3.4: All models are fitted using panel regression with firm and date fixed effects and clustered standard error of date. Robust t-statistics are reported in parentheses. ***,**,* indicate the significance level of 1%, 5%, and 10%, respectively. Key results are highlighted.

	RH		TAQ	
	(1)	(2)	(3)	(4)
VARIABLES	*Ret (t)*	*Ret (t)*	*Ret (t)*	*Ret (t)*
Retail Trading	0.0463	0.0473	0.102***	0.102***
	(1.359)	(1.388)	(8.144)	(8.142)
Reddit10	0.354		-0.0941	
	(0.643)		(-0.706)	
Retail Trading×Reddit10	-0.0720		-0.00288	
	(-0.977)		(-0.00695)	
WSB Attn	0.00522*	0.00513*	0.00679**	0.00671**
	(1.849)	(1.817)	(2.524)	(2.493)
WSB Attn×Reddit10	**-0.372***		**-0.457****	
	(-1.739)		(-2.201)	
RH100		0.891**		-0.0839
		(2.248)		(-1.337)
Retail Trading×RH100		**-0.149****		-0.0535
		(-2.395)		(-0.664)
Observations	611,664	611,664	784,376	784,376
R-squared	0.083	0.083	0.072	0.072
Date FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes

same interaction term is not significant for TAQ traders. Moreover, when looking at the relationship between *Retail Trading* and *Ret(t)*, TAQ trading is positively and significantly related to stock return (the corresponding coefficient is in columns (3) and (4), 0.102^{***} and 0.102^{***}, respectively), and the RH trading does not establish a significant relationship with stock return on day t.

To summarize, RH retail trading on the popular RH stocks relates to lower stock return and higher idiosyncratic risks. TAQ retail trading on the same group of stocks is associated with lower idiosyncratic risks. Reddit attention on popular Reddit stocks corresponds to higher firms' idiosyncratic risk and lower stock return.

3.5 Conclusion

This paper examines the two types of retail investors, Robinhood and TAQ traders, commonly used in the literature as proxies for U.S. retail traders. I start by outlining the similarities and differences between RH trading and TAQ retail trading, particularly, (1)Both TAQ and RH experienced a sharp increase when it comes to the retail traders' participation in the stock market, (2) RH ticker universe appears to be much smaller than TAQ's, and RH investors typically trade cheaper stocks, and (3) RH investors establish a positive and persistent relationship with Reddit WSB Attention, while TAQ traders relationship with Reddit WSB Attention is weakly negative. The first set of stylized facts provides evidence of the heterogeneity between RH and TAQ investors. It also confirms that the younger generation of retail investors rely more on social media compared to the more experienced generation. The differences between the two types could be one of the sources as to why there are heterogeneous findings in the literature regarding the role of retail traders.

I further assess the relationship between TAQ and RH net buying behavior. According to Pedersen (2021), fanatic investors would only listen to themselves and invest in stocks they prefer, and rational investors, on the other hand, listen to everyone and update their beliefs. This implies that RH net buying should have some predictive power of TAQ net buying. I am only interested in the relationship between RH and TAQ net buying on an average level,

and the findings suggest they are only remotely connected. Therefore, the main takeaway in this part is that TAQ traders, on average, do not respond to RH net buying. In other words, if there is a concentrated-trading episode for RH investors, it is unlikely that it will spill over to other retail traders.

Lastly, I focus on the popular stock tickers, either on Reddit WSB or RH. I then examine the relationship between idiosyncratic volatility (or stock return) and retail trading for those stocks. I create two lists, RH100 and Reddit10, which contain the daily top 100 most held stock on RH and the daily top 10 most discussed tickers on Reddit WSB, respectively. These lists may contain the top gainers/losers of the day and some meme stocks. Higher WSB Attention around Reddit 10 is associated with negative returns and higher idiosyncratic volatility. More intensive RH trading of the RH 100 stocks also relates to negative returns and higher idiosyncratic volatility. Notably, higher TAQ marketable retail order imbalances are associated with lower idiosyncratic risk. This finding suggests the market implications of RH and TAQ traders are different. From a return and idiosyncratic volatility point of view, RH traders are less informed than TAQ traders.

To summarize, this chapter aims to provide more stylized facts about the similarities and differences between RH and TAQ traders. It also confirms that RH and TAQ respond to information on social media differently, and when it comes to the relationship between retail trading and stock return/idiosyncratic volatility, TAQ and RH's differences only deepened. The findings imply TAQ traders are more informed than RH traders. There is some evidence in my findings that support the mechanism proposed in Pedersen (2021) on the interaction between rational and fanatic investors.

Chapter 4

CONCLUSION

The three chapters of my dissertation are closely related but also different in many ways. The first chapter focuses on the social discussion board Reddit $r/WallStreetBets$. I was often asked the reason that I choose Reddit and how it is different from, say, the news outlets, or other social media platforms. My answer is that Reddit $r/WallStreetBets$ has some unique characteristics that create lots of potential when it comes to retail trading. Different from a Google search index, and analysis on firm reports, Reddit, as an online social network, provides a sense of community, where users discuss, debate, and learn about the stock market. Moreover, because of the $GME short squeeze event, Reddit WSB has earned wide attention from the retail trader community, and also a reputation of irrational herding and reckless trading. These observations serve as motivations for me to look at Reddit more in depth, and also on its impact on the financial market, and more importantly, how WSB influencers, such as "Deep*******Value" affect how other users think about the market. I find Reddit WSB on average respond to most of the firm's key development events, which is a proxy for firms' fundamental information. This makes sense because several studies, using different periods of Reddit data and measurements, suggested Reddit WSB activities, to some degree, can predict or explain stock return. What's more comforting to me is that I also find WSB influencers on average cannot represent the entire community. Instead of thinking Reddit WSB as a "mutual fund" managed by a couple of influencers, I see the wisdom of the crowd in my analysis.

The second chapter aims to bridge the gap in the literature. Eaton et al. (2022) and Ozik et al. (2021) gave different results when it comes to the relationship between RH trading and volatility. We argue that they are both correct and their results complement each other.

The reason their results are not conflicting is that they essentially use different volatility. Eaton et al. (2022) calculates volatility as the daily average of returns for 5-minute intervals, which we show is basically idiosyncratic volatility. Ozik et al. (2021) measures volatility by calculating the daily standard deviation of returns, which we show reflects more of the total volatility of the stocks. We also use both theoretical and empirical models to link Reddit WSB sentiment and volatility. We show that WSB sentiment increases idiosyncratic volatility but decreases total volatility. This chapter uses one type of retail trading, the RH traders, and Reddit sentiment, to provide more insights on the role of retail traders in the financial market.

In the last chapter, I include one more type of retail trader, the TAQ traders. They are essentially retail traders who don't use Robinhood as their trading platform (under some conditions, described in the previous chapter). I am particularly interested in the interaction between the two types of traders, and how they differ in responding to the online social discussion boards (Reddit WSB in my case). The findings confirm my hypothesis that RH traders are less informed compared to TAQ investors. This provides a more clear picture of the generational shift that's happening with retail traders. Younger, inexperienced, less wealthy traders (RH investors) collect information and trade in a very different way from the richer and more experienced retail traders. These differences could stem from the user interface (UI) design of the trading platforms, different risk tolerance, information-gathering process,..., etc. With that being said, there are some limitations in my work and I describe some of them and potential future work in the next section.

4.1 Limitations and Future Work

My work around Reddit *r/WallStreetBets* only considers textual data. Although emojis or the like can be considered using VADER to calculate sentiment, memes are not included. As a Reddit WSB user, I am aware that memes can contain lots of information, and potentially contribute to aggregate attention and sentiment. Memes are also closely related to "meme stocks", which defines as the group of stocks that were frequently mentioned by memes. I

made an effort to include popular stocks on Reddit, for example, the top 10 Reddit stock list, however, more work can be done in evaluating the information conveyed in memes. This is particularly useful when it comes to understanding the "meme trading" behavior and predicting events like the GME short squeeze event.

The limitation of chapter 3 is probably generalizability. It only provides a snapshot, for a certain period of time, of how TAQ and RH traders behave differently. The notion of RH being less informed compared to TAQ traders is also subject to data availability. With the development of high-frequency traders, and more institutional investors placing small trading orders (just like the retail traders), the RH trades could result in the TAQ database eventually, so that we don't observe a separation of different types of retail traders.

On the final note, I had tremendous fun looking at the retail traders and Reddit. I believe that technology is blurring the line between the informed and the uninformed. Reddit opens a new door for retail traders and I can not wait to see how the new AI (for example, ChatGPT) changes the game. Could there be one day when everyone can be a skilled/informed trader?

BIBLIOGRAPHY

T. Aabo, C. Pantzalis, and J. C. Park. Idiosyncratic volatility: An indicator of noise trading? *Journal of Banking & Finance*, 75:136–151, 2017.

D. Adjodah, Y. Leng, S. K. Chong, P. Krafft, E. Moro, and A. Pentland. Accuracy-risk trade-off due to social learning in crowd-sourced financial predictions. *Entropy*, 23(7):801, 2021.

Y. Aït-Sahalia, P. A. Mykland, and L. Zhang. Ultra high frequency volatility estimation with dependent microstructure noise. *Journal of Econometrics*, 160(1):160–175, 2011.

Y. Amihud. Illiquidity and stock returns: cross-section and time-series effects. *Journal of financial markets*, 5(1):31–56, 2002.

A. Anand and J. Pathak. The role of reddit in the gamestop short squeeze. *Economics Letters*, 211:110249, 2022.

A. Ang, R. J. Hodrick, Y. Xing, and X. Zhang. High idiosyncratic volatility and low returns: International and further us evidence. *Journal of Financial Economics*, 91(1):1–23, 2009.

D. Ardia, C. Aymard, and T. Cenesizoglu. Fast and furious: An intraday analysis of robinhood users? trading behavior. *Available at SSRN*, 2022.

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

T. G. Bali, D. Hirshleifer, L. Peng, and Y. Tang. Attention, social interaction, and investor attraction to lottery stocks. Technical report, National Bureau of Economic Research, 2021.

R. Ball and P. Brown. An empirical evaluation of accounting income numbers. *Journal of accounting research*, pages 159–178, 1968.

B. M. Barber and T. Odean. Trading is hazardous to your wealth: The common stock investment performance of individual investors. *The journal of Finance*, 55(2):773–806, 2000.

B. M. Barber and T. Odean. The internet and the investor. *Journal of Economic Perspectives*, 15(1):41–54, 2001.

B. M. Barber and T. Odean. 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, 2008.

B. M. Barber, X. Huang, T. Odean, and C. Schwarz. Attention induced trading and returns: Evidence from robinhood users. *Journal of Finance, Forthcoming*, 2021.

S. Behrendt and A. Schmidt. The twitter myth revisited: Intraday investor sentiment, twitter activity and individual-level stock return volatility. *Journal of Banking & Finance*, 96: 355–367, 2018.

G. Bekaert, R. J. Hodrick, and X. Zhang. Aggregate idiosyncratic volatility. *Journal of Financial and Quantitative Analysis*, 47(6):1155–1185, 2012.

S. Bhagat, M. W. Marr, and G. R. Thompson. The rule 415 experiment: Equity markets. *The Journal of Finance*, 40(5):1385–1401, 1985.

E. Boehmer and W. Song. Smart retail traders, short sellers, and stock returns. *Short Sellers, and Stock Returns (October 23, 2020)*, 2020.

E. Boehmer, C. M. Jones, X. Zhang, and X. Zhang. Tracking retail investor activity. *The Journal of Finance*, 76(5):2249–2305, 2021.

S. Borovkova and D. Mahakena. News, volatility and jumps: the case of natural gas futures. *Quantitative Finance*, 15(7):1217–1242, 2015.

D. Bradley, J. Hanousek Jr, R. Jame, and Z. Xiao. Place your bets? the market consequences of investment research on reddit's wallstreetbets. *The Market Consequences of Investment Research on Reddit's Wallstreetbets (March 15, 2021)*, 2021.

B. Bushee, M. Cedergren, and J. Michels. Does the media help or hurt retail investors during the ipo quiet period? *Journal of Accounting and Economics*, 69(1):101261, 2020.

D. Cahill, M. Wee, and J. W. Yang. Media sentiment and trading strategies of different types of traders. *Pacific-Basin Finance Journal*, 44:160–172, 2017.

J. Y. Campbell and A. S. Kyle. Smart money, noise trading and stock price behaviour. *The Review of Economic Studies*, 60(1):1–34, 1993.

R. G. Chacon, T. G. Morillon, and R. Wang. Will the reddit rebellion take you to the moon? evidence from wallstreetbets. *Financial Markets and Portfolio Management*, pages 1–25, 2022.

K. Chan, K. C. Chan, and G. A. Karolyi. Intraday volatility in the stock index and stock index futures markets. *The Review of Financial Studies*, 4(4):657–684, 1991.

I. Cioroianu, S. Corbet, and C. Larkin. The differential impact of corporate blockchain-development as conditioned by sentiment and financial desperation. *Journal of Corporate Finance*, 66:101814, 2021.

J. A. Cookson, J. E. Engelberg, and W. Mullins. Echo chambers. *The Review of Financial Studies*, 36(2):450–500, 2023.

Z. Da, V. W. Fang, and W. Lin. Fractional trading. *Available at SSRN 3949697*, 2021.

J. B. De Long, A. Shleifer, L. H. Summers, and R. J. Waldmann. Noise trader risk in financial markets. *Journal of political Economy*, 98(4):703–738, 1990.

M. H. DeGroot. Reaching a consensus. *Journal of the American Statistical association*, 69 (345):118–121, 1974.

P. M. DeMarzo, D. Vayanos, and J. Zwiebel. Persuasion bias, social influence, and unidimensional opinions. *The Quarterly journal of economics*, 118(3):909–968, 2003.

D. J. Denis and A. Sarin. Ownership and board structures in publicly traded corporations. *Journal of Financial Economics*, 52(2):187–223, 1999.

G. W. Eaton, T. C. Green, B. S. Roseman, and Y. Wu. Retail trader sophistication and stock market quality: Evidence from brokerage outages. *Journal of Financial Economics*, 146(2):502–528, 2022.

L. G. Epstein and M. Schneider. Ambiguity, information quality, and asset pricing. *The Journal of Finance*, 63(1):197–228, 2008.

E. F. Fama and K. R. French. The cross-section of expected stock returns. *the Journal of Finance*, 47(2):427–465, 1992.

M. Farrell, T. C. Green, R. Jame, and S. Markov. The democratization of investment research and the informativeness of retail investor trading. *Journal of Financial Economics*, 145 (2):616–641, 2022.

A. R. Gallant, P. E. Rossi, and G. Tauchen. Stock prices and volume. *The Review of Financial Studies*, 5(2):199–242, 1992.

M. S. Gerety and J. H. Mulherin. Patterns in intraday stock market volatility, past and present. *Financial Analysts Journal*, 47(5):71–79, 1991.

R. Greenwood, T. Laarits, and J. Wurgler. Stock market stimulus. Technical report, National Bureau of Economic Research, 2022.

T. Hasso, D. Müller, M. Pelster, and S. Warkulat. Who participated in the gamestop frenzy? evidence from brokerage accounts. *Finance Research Letters*, 45:102140, 2022.

D. Hu, C. M. Jones, V. Zhang, and X. Zhang. The rise of reddit: How social media affects retail investors and short-sellers? roles in price discovery. *Available at SSRN 3807655*, 2021.

C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.

N. Jayaram. The rising power of the individual investor: How social media sentiments and user activity impact stock price volatility and trading volume. 2022.

K. John and J. Li. Covid-19, volatility dynamics, and sentiment trading. *Journal of Banking & Finance*, 133:106162, 2021.

K. Joseph, M. B. Wintoki, and Z. Zhang. Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search. *International Journal of Forecasting*, 27(4):1116–1127, 2011.

B. Ke and K. Petroni. How informed are actively trading institutional investors? evidence from their trading behavior before a break in a string of consecutive earnings increases. *Journal of Accounting Research*, 42(5):895–927, 2004.

P. J. Kelly. Information efficiency and firm-specific return variation. *The Quarterly Journal of Finance*, 4(04):1450018, 2014.

B. Kőszegi and F. Matějka. Choice simplification: A theory of mental budgeting and naive diversification. *The Quarterly Journal of Economics*, 135(2):1153–1207, 2020.

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

S. Laurent and S. Shi. Volatility estimation and jump detection for drift–diffusion processes. *Journal of Econometrics*, 217(2):259–290, 2020.

B.-S. Lee and O. M. Rui. The dynamic relationship between stock returns and trading volume: Domestic and cross-country evidence. *Journal of Banking & Finance*, 26(1): 51–78, 2002.

W. Y. Lee, C. X. Jiang, and D. C. Indro. Stock market volatility, excess returns, and the role of investor sentiment. *Journal of banking & Finance*, 26(12):2277–2299, 2002.

S. Lin, P. F. Pope, and S. Young. Stock market reaction to the appointment of outside directors. *Journal of Business Finance & Accounting*, 30(3-4):351–382, 2003.

G. Llorente, R. Michaely, G. Saar, and J. Wang. Dynamic volume-return relation of individual stocks. *The Review of financial studies*, 15(4):1005–1047, 2002.

L. J. Lockwood and S. C. Linn. An examination of stock market return volatility during overnight and intraday periods, 1964–1989. *The Journal of Finance*, 45(2):591–601, 1990.

S. Long, B. Lucey, Y. Xie, and L. Yarovaya. "i just like the stock": The role of reddit sentiment in the gamestop share rally. *Financial Review*, 58(1):19–37, 2023.

T. Loughran and B. McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65, 2011.

T. Loughran and B. McDonald. Textual analysis in finance. *Annual Review of Financial Economics*, 12:357–375, 2020.

A. M. Malz. The gamestop episode: What happened and what does it mean? *Journal of Applied Corporate Finance*, 33(4):87–97, 2021.

B. Mendel and A. Shleifer. Chasing noise. *Journal of Financial Economics*, 104(2):303–320, 2012.

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

M. S. Pagano, J. Sedunov, and R. Velthuis. How did retail investors respond to the covid-19 pandemic? the effect of robinhood brokerage customers on market quality. *Finance Research Letters*, 43:101946, 2021.

L. H. Pedersen. Game on: Social networks and markets. *Available at SSRN 3794616*, 2021.

J. Pontiff. Costly arbitrage and the myth of idiosyncratic risk. *Journal of Accounting and Economics*, 42(1-2):35–52, 2006.

D. Rakowski, S. E. Shirley, and J. R. Stark. Twitter activity, investor attention, and the diffusion of information. *Financial Management*, 50(1):3–46, 2021.

R. Roll. R-squared. *Journal of finance*, 43(2):541–566, 1988.

J. Sahlberg. Online herding activity. 2021.

A. Sanford. Does perception matter in asset pricing? modeling volatility jumps using twitter-based sentiment indices. *Journal of Behavioral Finance*, 23(3):262–280, 2022.

P. G. Savor. Stock returns after major price shocks: The impact of information. *Journal of financial Economics*, 106(3):635–659, 2012.

S. W. Seo and J. S. Kim. The information content of option-implied information for volatility forecasting with investor sentiment. *Journal of Banking & Finance*, 50:106–120, 2015.

W. F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442, 1964.

R. J. Shiller. *Market Volatility*. MIT Press, Cambridge, MA, 1989.

A. Shleifer and R. W. Vishny. A survey of corporate governance. *The journal of finance*, 52 (2):737–783, 1997.

I. Welch. The wisdom of the robinhood crowd. Technical report, National Bureau of Economic Research, 2020.

I. Welch. The wisdom of the robinhood crowd. *The Journal of Finance*, 77(3):1489–1527, 2022.

J. Wong. The impact of subreddit comments on daily return and volume. 2021.

Y. C. Yang, B. Zhang, and C. Zhang. Is information risk priced? evidence from abnormal idiosyncratic volatility. *Journal of Financial Economics*, 135(2):528–554, 2020.

I. Yousaf, L. Pham, and J. W. Goodell. The connectedness between meme tokens, meme stocks, and other asset classes: Evidence from a quantile connectedness approach. *Journal of International Financial Markets, Institutions and Money*, 82:101694, 2023.

Appendix A

APPENDIX

A.1 Alternative Model on Idiosyncratic Volatility and Total Volatility

Our goal in this model is to provide mathematical intuition for the mechanism described in our paper. In particular, we will show that as the time interval for the estimation of the standard deviation of returns approaches zero, that the estimated volatility reflects, in large part, the idiosyncratic component of risk. We start by assuming a data-generating process using a geometric brown motion:

$$dX_t = \mu_t dt + \sigma_t dW_t \tag{A.1}$$

We begin, as others have done (Aït-Sahalia et al., 2011), by assuming that at very short intervals, the drift for a geometric Brownian motion is equal to zero. In other words, under the condition that the Δ of time is smaller (e.g. seconds), the drift is irrelevant, both economically and statistically. Thus, we focus on the function of the σ_t and allow $\mu_t = 0$. We then assume that the total volatility can be decomposed into a systematic and an idiosyncratic component, such that:

$$\sigma(t) = \sigma_{sys}(t) + \sigma_{idio}(t) \tag{A.2}$$

where $\sigma_{sys}(t)$ is the systematic volatility at time t and $\sigma_{idio}(t)$ is the idiosyncratic volatility at time t. Again, our goal is to show that, as the time interval in the calculation of the volatility approaches zero, the proportion of total volatility to idiosyncratic volatility approaches one, in other words, the share of total volatility represented by systematic volatility becomes zero at the limit. Volatility in the short term is, in essence, mostly capturing noise (Behrendt and Schmidt, 2018; Gerety and Mulherin, 1991; Lockwood and Linn, 1990). Fundamental

or systematic changes cause jumps in the volatility process but, for example, at the seconds interval, these are few and far between. To calculate the proportion of total volatility represented by the idiosyncratic volatility, we need to compute the variance of the idiosyncratic volatility, denoted as $Var(\sigma_{idio}(t))$, and the variance of the total volatility denoted as $Var(\sigma(t))$. The variance of the idiosyncratic volatility can be expressed as:

$$Var(\sigma_{idio}(t)) = E[(\sigma_{idio}(t) - E[\sigma_{idio}(t)])^2] \tag{A.3}$$

where $E[\cdot]$ denotes the expected value. Similarly, the variance of the total volatility can be expressed as:

$$Var(\sigma(t)) = E[(\sigma(t) - E[\sigma(t)])^2] \tag{A.4}$$

Now, we consider the limit as the time interval for the volatility, denoted as Δ_t, approaches zero:

$$\lim_{\Delta t \to 0} \frac{Var(\sigma_{idio}(t))}{Var(\sigma(t))} \tag{A.5}$$

Using Ito's lemma and some algebraic manipulations, it can be shown that:

$$Var(\sigma_{idio}(t)) = \int_0^t E[\sigma_{idio}^2(u)]du$$
$$Var(\sigma(t)) = \int_0^t E[\sigma^2(u)]du$$

where the integrals are taken over the time interval $[0, t]$. By the law of large numbers, we know that:

$$\lim_{\Delta t \to 0} \frac{\int_0^t E[\sigma_{idio}^2(u)]du}{\Delta t} = E[\sigma_{idio}^2(t)]$$
$$\lim_{\Delta t \to 0} \frac{\int_0^t E[\sigma^2(u)]du}{\Delta t} = E[\sigma^2(t)]$$

Therefore, we have:

$$\lim_{\Delta t \to 0} \frac{Var(\sigma_{idio}(t))}{Var(\sigma(t))} = \frac{E[\sigma_{idio}^2(t)]}{E[\sigma^2(t)]} \tag{A.6}$$

Since both $\sigma_{sys}(t)$ and $\sigma_{idio}(t)$ are non-negative, we have:

$$\sigma^2(t) = (\sigma_{sys}(t) + \sigma_{idio}(t))^2 \geq \sigma_{idio}^2(t) \tag{A.7}$$

Therefore, the proportion of total volatility represented by the idiosyncratic volatility is always less than or equal to one. Moreover, since the idiosyncratic volatility is uncorrelated with the systematic volatility, we have:

$$E[\sigma^2_{idio}(t)] \geq 0$$
$$E[\sigma^2(t)] > 0$$

Therefore, as the time interval for the volatility approaches zero, we have:

$$\lim_{\Delta t \to 0} \frac{Var(\sigma_{idio}(t))}{Var(\sigma(t))} \geq 0 \tag{A.8}$$

This then ensures that the ratio is at least greater than or equal to zero. Using L'Hopital's rule, we can take the derivative of both the numerator and denominator of the ratio as the time interval approaches zero, and then evaluate the limit of the resulting ratio. Specifically, we have:

$$\begin{aligned}\lim_{\Delta t \to 0} \frac{Var(\sigma_{idio}(t))}{Var(\sigma(t))} &= \lim_{\Delta t \to 0} \frac{\int_0^t E[\sigma^2_{idio}(u)]du}{\int_0^t E[\sigma^2(u)]du} \\ &= \lim_{\Delta t \to 0} \frac{E[\sigma^2_{idio}(t)]}{E[\sigma^2(t)]} \\ &= \frac{E[\sigma^2_{idio}(t)]}{E[\sigma^2(t)]}\end{aligned}$$

where we used the same steps as before to get to the second line, and the third line follows from the fact that the expected values of the idiosyncratic and total volatility are continuous functions of time, and therefore the limit of their ratio as the time interval approaches zero is simply the ratio of their values at time t. Now, to evaluate the limit of the ratio $E[\sigma^2_{idio}(t)]/E[\sigma^2(t)]$, we can take the derivative of both the numerator and denominator with respect to time t using the chain rule:

$$\frac{d}{dt}E[\sigma^2_{idio}(t)] = 2\sigma_{idio}(t)E[\sigma'_{idio}(t)]$$
$$\frac{d}{dt}E[\sigma^2(t)] = 2\sigma(t)E[\sigma'(t)]$$

where $\sigma'_{idio}(t)$ and $\sigma'(t)$ denote the time derivatives of the idiosyncratic and total volatility, respectively. Now, as the time interval approaches zero, we have:

$$\begin{aligned}\lim_{\Delta t \to 0} \frac{E[\sigma^2_{idio}(t)]}{E[\sigma^2(t)]} &= \lim_{\Delta t \to 0} \frac{2\sigma_{idio}(t)E[\sigma'_{idio}(t)]}{2\sigma(t)E[\sigma'(t)]} \\ &= \frac{\sigma_{idio}(t)}{\sigma(t)} \cdot \frac{E[\sigma'_{idio}(t)]}{E[\sigma'(t)]}\end{aligned}$$

where the second line follows from the fact that the constant factor of 2 in the numerator and denominator cancels out, and we can take the limit of the resulting ratio since both $\sigma_{idio}(t)$ and $\sigma(t)$ are positive and non-zero.

Now, since $\sigma_{idio}(t)$ is the idiosyncratic component of the total volatility, it represents the portion of the volatility that is specific to the individual stock, and is therefore more likely to be volatile and unpredictable compared to the systematic component $\sigma_{sys}(t)$. In other words, as the time interval approaches zero, the idiosyncratic volatility will likely fluctuate more rapidly than the systematic volatility, and this will cause the ratio $\sigma_{idio}(t)/\sigma(t)$ to increase. Furthermore, since the ratio $E[\sigma'_{idio}(t)]/E[\sigma'(t)]$ is bounded and continuous with respect to time, we can conclude that the limit of the ratio $E[\sigma^2_{idio}(t)]/E[\sigma^2(t)]$ as the time interval approaches zero is greater than or equal to zero, and therefore the proportion of total volatility represented by the idiosyncratic volatility gets larger and larger as the time interval approaches zero. Thus, using L'Hopital's rule and the properties of the idiosyncratic and total volatility, we have shown mathematically that the proportion of total volatility represented by the idiosyncratic volatility actually gets larger as the time interval approaches zero.

So, how do we know that what we showed above is not also the case for the systematic risk component? To complete our proof, we also need to argue that the opposite is not true. By construction, this will need to be the case just by the property of complements. In essence, since total risk was decomposed into two components, $\sigma(t) = \sigma_{sys}(t) + \sigma_{idio}(t)$, and the ratio of idiosyncratic risk to total risk is getting larger, by definition that implies that the ratio of systematic risk to total risk will be getting smaller, all else equal, as the time interval approaches zero. A less mathematical explanation for this phenomenon, however, is that as the time interval approaches zero, while the proportion of total volatility represented by the

systematic volatility gets smaller and smaller as the time interval approaches zero, is because of the relative magnitudes of the two components. Recall that the systematic volatility is the component of volatility that is common to all stocks, and is driven by macroeconomic and other external factors that affect the entire market. On the other hand, the idiosyncratic volatility is the component of volatility that is specific to individual stocks, and is driven by company-specific factors such as earnings surprises or changes in management. Since the systematic volatility is common to all stocks, it tends to be less volatile and more predictable than the idiosyncratic volatility. As a result, as the time interval approaches zero, the idiosyncratic volatility will fluctuate more rapidly than the systematic volatility, and this will cause the ratio of idiosyncratic volatility to total volatility to increase, while the ratio of systematic volatility to total volatility will decrease.

In other words, the proportion of total volatility represented by the idiosyncratic volatility gets larger as the time interval approaches zero because the idiosyncratic volatility is more volatile and unpredictable than the systematic volatility, while the proportion of total volatility represented by the systematic volatility gets smaller as the time interval approaches zero because the systematic volatility is less volatile and more predictable than the idiosyncratic volatility.

A.2 Tables

The following pages include some regressions, top 10 Key Development events, and lists of popular stocks on Robinhood or Reddit over 2018 to 2021. I also include a sample of Reddit tickers that were removed from my dataset here, for example, “DD”, “CEO”, “YOLO”, “IPO”, “MEME”, “BUY”, “SELL”, “GDP”, “FLY”, “NEW”, “BABY”, “USD”, “UK”, “LOVE”, “LIFE”, “CPI”].

Table A.1: Top 10 Key Development Events

Top 10 Key Developments		
	Name	Description
1	Conferences	Events hosted by investment companies where various companies participate to update analysts, shareholders, investors etc. regarding issues related to company.
2	Company Conference Presentations	This pertains to when a company participates in a "Conference Call" conducted by third parties, to discuss on any issue with its Shareholders, Analysts and Investors.
3	Earning Announcement	Announcements of quarterly, annual, or other periodic earnings.
4	Product Related Announcements	Announcements pertaining to the introduction, change, improvement, or discontinuation of a company's product or services.
5	Earning Calls	This pertains to when a company conducts a "Conference Call" to discuss its quarterly, annual, or other periodic earnings.
6	Earning Release Date	Future dates when Earnings are going to be announced.
7	Executive Board Changes	Either a change in the membership of an elected board of directors whose members serve terms or changes in executive level positions within a firm, excluding CEO and CFO changes.

Top 10 Key Development Events		
	Name	Description
8	Client Announcements	An announcement of the beginning, ending or change in a relationships between a corporation and their clients or potential future clients.
9	Buyback Tranche Update	In a buyback transaction, the company plans to repurchase a certain amount of its own shares. The company discloses the repurchase information either for a quarter/half year/annual or for random periods falling between the duration of the buyback plan. This repurchases information (i.e. how much actually the company has bought back during a particular period (within the entire duration of the plan) is what S&P Global Market Intelligence calls tranches.
10	Corporate Guidance	An announcement of the expectation of operating results. It can be lower, remains confident, higher than its previously annouced expectation.

Table A.2: Panel regressions with Reddit attention based on all submissions and posts as dependent. The independent variables are firms' key development events. Date and Firm fixed effects are included. The regression is clustered by date.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	*Reddit_Attn_All*	*Reddit_Attn_All*	*Reddit_Attn_All*	*Reddit_Attn_All*	*Reddit_Attn_All*	*Reddit_Attn_All*	*Reddit_Attn_All*	*Reddit_Attn_All*	*Reddit_Attn_All*	*Reddit_Attn_All*
Conferences	-0.172									
	(-1.036)									
Company_Conference_Presentations		-0.351**								
		(-1.962)								
Earning_Announcement			0.161							
			(0.810)							
Product_Related_Announcements				-0.525*						
				(-1.700)						
Earning_Calls					-0.543*					
					(-1.863)					
Earning_Release_Date						-0.578*				
						(-1.866)				
Executive_Board_Changes							-0.546**			
							(-2.357)			
Client_Announcements								-0.102		
								(-0.204)		
Buyback_Tranche_Update									0.277	
									(0.835)	
Corporate_Guidance										0.122
										(0.404)
Observations	3,219,258	3,219,258	3,219,258	3,219,258	3,219,258	3,219,258	3,219,258	3,219,258	3,219,258	3,219,258
R-squared	0.046	0.046	0.046	0.046	0.046	0.046	0.046	0.046	0.046	0.046
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table A.3: Panel regressions with Reddit VADER sentiments based on all submissions and posts as dependent. The independent variables are firms' key development events. Date and Firm fixed effects are included. The regression is clustered by date.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	*Reddit_Sent_VADER*	*Reddit_Sent_VADER*	*Reddit_Sent_VADER*	*Reddit_Sent_VADER*	*Reddit_Sent_VADER*	*Reddit_Sent_VADER*	*Reddit_Sent_VADER*	*Reddit_Sent_VADER*	*Reddit_Sent_VADER*	*Reddit_Sent_VADER*
Conferences	-0.172									
	(-1.036)									
Company_Conference_Presentations		-0.351**								
		(-1.962)								
Earning_Announcement			0.161							
			(0.810)							
Product_Related_Announcements				-0.525*						
				(-1.700)						
Earning_Calls					-0.543*					
					(-1.863)					
Earning_Release_Date						-0.578*				
						(-1.866)				
Executive_Board_Changes							-0.546**			
							(-2.357)			
Client_Announcements								-0.102		
								(-0.204)		
Buyback_Tranche_Update									0.277	
									(0.835)	
Corporate_Guidance										0.122
										(0.404)
Observations	3,219,258	3,219,258	3,219,258	3,219,258	3,219,258	3,219,258	3,219,258	3,219,258	3,219,258	3,219,258
R-squared	0.046	0.046	0.046	0.046	0.046	0.046	0.046	0.046	0.046	0.046
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ticker FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table A.4: Top 30 Robinhood Tickers

No.	**Ticker**	**RHH**	**Average Price**	**Type**
1	ACB	460,731	$ 6.39	Common
2	F	344,576	$ 8.89	Common
3	GE	327,691	$ 10.32	Common
4	AAPL	246,628	$ 219.17	Common
5	MSFT	241,720	$ 141.60	Common
6	GPRO	226,621	$ 5.29	Common
7	FIT	205,573	$ 5.78	Common
8	DIS	197,600	$ 121.87	Common
9	AMD	170,817	$ 37.66	Common
10	CRON	166,894	$ 9.52	Common
11	SNAP	164,227	$ 16.44	Common
12	NKLA	158,919	$ 34.30	Common
13	FB	157,022	$ 195.98	Common
14	HEXO	155,390	$ 2.76	Common
15	TOPS	154,607	$ 1.04	Common
16	TSLA	153,656	$ 459.83	Common
17	UBER	146,529	$ 35.52	Common
18	CGC	142,008	$ 28.73	Common
19	BAC	134,607	$ 28.52	Common
20	PLUG	134,463	$ 5.01	Common

Continued on next page

Table A.4: Top 30 Robinhood Tickers (Continued)

No.	Ticker	RHH	Average Price	Type
21	AMZN	130,372	$ 2,038.63	Common
22	TWTR	126,461	$ 35.21	Common
23	BABA	124,022	$ 197.06	ADRs
24	NFLX	108,813	$ 365.16	Common
25	ZNGA	103,854	$ 5.96	Common
26	AAL	100,755	$ 29.62	Common
27	NIO	100,617	$ 9.51	ADRs
28	DAL	96,693	$ 48.23	Common
29	NVDA	95,262	$ 267.75	Common
30	SBUX	91,264	$ 73.97	Common

Table A.5: Top 30 TAQ Tickers

No.	Ticker	Total Retail Trades	Average Price	Type
1	AAPL	75,232,178	$ 219.17	Common
2	TSLA	66,616,491	$ 459.83	Common
3	AMC	52,473,377	$ 10.84	Common
4	AMD	47,203,505	$ 37.66	Common
5	NIO	39,184,049	$ 9.51	ADRs
6	NVDA	33,026,568	$ 267.75	Common

Continued on next page

Table A.5: Top 30 TAQ Tickers (Continued)

No.	**Ticker**	**Total Retail Trades**	**Average Price**	**Type**
7	MSFT	29,760,654	$ 141.60	Common
8	TQQQ	29,082,697	$ 89.46	Units
9	SPY	28,520,868	$ 295.58	Units
10	AMZN	27,949,217	$ 2,038.63	Common
11	SNDL	24,419,733	$ 2.19	Common
12	FB	23,207,571	$ 195.98	Common
13	SQQQ	22,833,233	$ 18.42	Units
14	QQQ	22,515,110	$ 202.41	Units
15	F	20,849,283	$ 8.89	Common
16	BA	20,562,352	$ 302.08	Common
17	BABA	18,892,881	$ 197.06	ADRs
18	GE	16,370,150	$ 10.32	Common
19	DIS	16,012,852	$ 121.87	Common
20	GME	15,978,704	$ 9.83	Common
21	T	15,669,293	$ 32.77	Common
22	AAL	15,496,339	$ 29.62	Common
23	PLTR	15,489,992	$ 17.76	Common
24	BAC	14,502,120	$ 28.52	Common
25	NFLX	13,924,790	$ 365.16	Common

Continued on next page

Table A.5: Top 30 TAQ Tickers (Continued)

No.	**Ticker**	**Total Retail Trades**	**Average Price**	**Type**
26	CCL	13,871,164	$ 44.19	Units
27	MRNA	13,491,655	$ 39.62	Common
28	GOOG	13,213,477	$ 1,261.49	Common
29	PFE	12,769,465	$ 38.56	Common
30	SQ	12,723,008	$ 83.96	Common

Table A.6: Top 30 Reddit Tickers

No.	**Ticker**	**# of Mentions**	**Average Price**	**Type**
1	TSLA	73,963	$ 459.83	Common
2	SPY	65,142	$ 295.58	Units
3	PLTR	52,271	$ 17.76	Common
4	AMD	42,585	$ 37.66	Common
5	GME	40,524	$ 9.83	Common
6	MU	27,026	$ 47.36	Common
7	MSFT	20,338	$ 141.60	Common
8	RH	19,843	$ 178.16	Common
9	F	19,827	$ 8.89	Common
10	NIO	15,933	$ 9.51	ADRs

Continued on next page

Table A.6: Top 30 Reddit Tickers (Continued)

No.	Ticker	# of Mentions	Average Price	Type
11	AAPL	14,571	$ 219.17	Common
12	AMZN	14,024	$ 2,038.63	Common
13	BYND	13,842	$ 125.24	Common
14	SPCE	13,223	$ 17.94	Common
15	FB	11,563	$ 195.98	Common
16	SP	11,562	$ 32.26	Common
17	DIS	10,370	$ 121.87	Common
18	GE	9,362	$ 10.32	Common
19	NKLA	8,852	$ 34.30	Common
20	SNAP	8,790	$ 16.44	Common
21	ATH	8,707	$ 41.82	Common
22	NVDA	7,389	$ 267.75	Common
23	EPS	7,387	$ 32.96	Units
24	PS	7,054	$ 22.37	Common
25	ITM	6,999	$ 42.61	Units
26	RSI	6,878	$ 21.58	Common
27	NFLX	6,824	$ 365.16	Common
28	PRPL	6,326	$ 11.06	Common
29	BA	6,114	$ 302.08	Common

Continued on next page

Table A.6: Top 30 Reddit Tickers (Continued)

No.	Ticker	# of Mentions	Average Price	Type
30	TA	6,023	$ 10.06	Common

Table A.7: RH Trading and Reddit Momentum

Panel A description						
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	RH Users Holding	RH Users Holding	RH Users Holding	RH Users Holding	RH Users Holding	RH Users Holding
Reddit Mom (t)	0.00856***	0.00568***	0.00523***	0.00471***	0.00389***	0.00312***
	(16.99)	(8.199)	(7.139)	(5.829)	(4.528)	(3.329)
Reddit Mom (t-1)		0.00501***	0.00363***	0.00365***	0.00351***	0.00321***
		(6.895)	(4.081)	(3.854)	(3.385)	(2.985)
Reddit Mom (t-2)			0.00307***	0.00168*	0.00207*	0.00232*
			(3.596)	(1.685)	(1.875)	(1.933)
Reddit Mom (t-3)				0.00352***	0.00161	0.00209*
				(3.686)	(1.537)	(1.957)
Reddit Mom (t-4)					0.00454***	0.00309***
					(4.396)	(2.614)
Reddit Mom (t-5)						0.00353***
						(3.873)
Date FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	809,216	767,760	719,002	670,415	614,412	563,950
R-squared	0.003	0.003	0.003	0.003	0.003	0.003

Table A.8: TAQ Trading and Reddit Momentum

Panel B						
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	mroib	mroib	mroib	mroib	mroib	mroib
Reddit Mom (t)	-0.000213	0.000163	8.09e-05	5.79e-05	0.000144	0.000400
	(-1.073)	(0.622)	(0.304)	(0.199)	(0.467)	(1.258)
Reddit Mom (t-1)		-0.000697***	-0.000916***	-0.000876***	-0.000899***	-0.000848**
		(-2.613)	(-2.977)	(-2.778)	(-2.666)	(-2.401)
Reddit Mom (t-2)			0.000379	0.000199	0.000139	-0.000112
			(1.084)	(0.524)	(0.349)	(-0.251)
Reddit Mom (t-3)				0.000289	0.000266	0.000238
				(0.790)	(0.632)	(0.553)
Reddit Mom (t-4)					6.53e-06	4.71e-05
					(0.0160)	(0.106)
Reddit Mom (t-5)						-0.000556*
						(-1.780)
Date FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	530,351	500,062	461,827	421,789	378,602	339,490
R-squared	0.002	0.002	0.002	0.002	0.002	0.002

Table A.9: This table presents how daily stock returns relate to Reddit activities and retail trading.

Panel A: Robinhood Traders

	(1)	(2)	(3)	(4)
VARIABLES	*Ret (t)*	*Ret (t)*	*Ret (t)*	*Ret (t)*
Users' Holding (Log)	0.0463	0.366**	0.0473	0.365**
	(1.359)	(2.203)	(1.388)	(2.199)
Reddit10	0.354	11.12		
	(0.643)	(1.469)		
Users' Holding × Reddit10	-0.0720	-0.949		
	(-0.977)	(-1.259)		
Reddit Mom	0.00522*		0.00513*	
	(1.849)		(1.817)	
Reddit Mom × Reddit10	**-0.372***			
	(-1.739)			
RH100			0.891**	0.195
			(2.248)	(0.0828)
RH100 × Users' Holding			**-0.149****	-0.0538
			(-2.395)	(-0.267)
Reddit Sent		0.123		0.128
		(1.392)		(1.437)
Reddit Sent × Reddit10		-12.91		
		(-1.644)		
Reddit Sent × RH100				-0.720
				(-1.210)
Observations	611,664	51,504	611,664	51,504
R-squared	0.083	0.230	0.083	0.230
Controls	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes

Table A.10: This table presents how daily stock returns relate to Reddit activities and retail trading.

Panel B: General Retail Traders				
	(1)	(2)	(3)	(4)
VARIABLES	*Ret (t)*	*Ret (t)*	*Ret (t)*	*Ret (t)*
mroib	0.102***	0.714***	0.102***	0.718***
	(8.144)	(5.852)	(8.142)	(5.851)
Reddit10	-0.0941	2.904*		
	(-0.706)	(1.794)		
mroib × *Reddit10*	-0.00288	2.236		
	(-0.00695)	(0.461)		
Reddit Mom	0.00679**		0.00671**	
	(2.524)		(2.493)	
Reddit Mom × *Reddit10*	**-0.457****			
	(-2.201)			
RH100			-0.0839	0.205
			(-1.337)	(0.379)
RH100 × *mroib*			-0.0535	-0.523
			(-0.664)	(-0.857)
Reddit Sent		0.0631		0.0680
		(0.904)		(0.968)
Reddit Sent × *Reddit10*		-4.966		
		(-1.271)		
Reddit Sent × *RH100*				-0.778
				(-1.531)
Observations	784,376	77,800	784,376	77,800
R-squared	0.072	0.157	0.072	0.157
Time FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes

Table A.11: This table presents how stock idiosyncratic volatility relates to Reddit activities and retail trading.

Panel A

	(1)	(2)	(3)	(4)
VARIABLES	*IVOL (t)*	*IVOL (t)*	*IVOL (t)*	*IVOL (t)*
Users' Holding	0.00816**	-0.0456***	0.00574	-0.0485***
	(1.977)	(-3.471)	(1.390)	(-3.708)
Reddit10	0.154	0.172		
	(0.872)	(0.276)		
Reddit10 × Users' Holding	-0.0248	-0.00733		
	(-0.984)	(-0.0953)		
Reddit Mom	**0.00148*** **		0.00148***	
	(3.450)		(3.470)	
Reddit10 × Reddit Mom	**0.239*** **			
	(3.157)			
Reddit Sent		0.0140		0.0143
		(0.912)		(0.918)
Reddit10 × Reddit Sent		0.365		
		(0.915)		
RH100			**-1.410*** **	**-0.965** **
			(-6.384)	(-2.356)
RH100 × Users' Holding			**0.253*** **	**0.0935** **
			(7.331)	(2.112)
RH100 × Reddit Mom			-0.0323*	
			(-1.742)	
RH100 × Reddit Sent				-0.0147
				(-0.102)
Observations	476,819	39,539	476,819	39,539
R-squared	0.169	0.420	0.169	0.420
Date FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes

Table A.12: This table presents how stock idiosyncratic volatility relates to Reddit activities and retail trading.

Panel B

	(1)	(2)	(3)	(4)
VARIABLES	*IVOL (t)*	*IVOL (t)*	*IVOL (t)*	*IVOL (t)*
mroib	-0.000600	0.00606	0.000594	0.00828
	(-0.185)	(0.311)	(0.180)	(0.425)
Reddit10	-0.0311	0.132		
	(-0.750)	(1.207)		
Reddit10 × mroib	0.0464	0.435		
	(0.291)	(0.782)		
Reddit Mom	**0.00168*** **		0.00168***	
	(4.130)		(4.146)	
Reddit10 × Reddit Mom	**0.217*** **			
	(3.656)			
Reddit Sent		0.00302		0.00358
		(0.264)		(0.311)
Reddit10 × Reddit Sent		-4.18e-05		
		(-0.000125)		
RH100			**0.181*** **	0.0317
			(9.787)	(0.394)
RH100 × mroib			**-0.107*** **	-0.239
			(-2.636)	(-0.907)
RH100 × Reddit Mom			-0.0125	
			(-0.541)	
RH100 × Reddit Sent				-0.0694
				(-0.520)
Observations	627,114	60,720	627,114	60,720
R-squared	0.148	0.367	0.148	0.367
Time FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes

A.3 Figures

The following

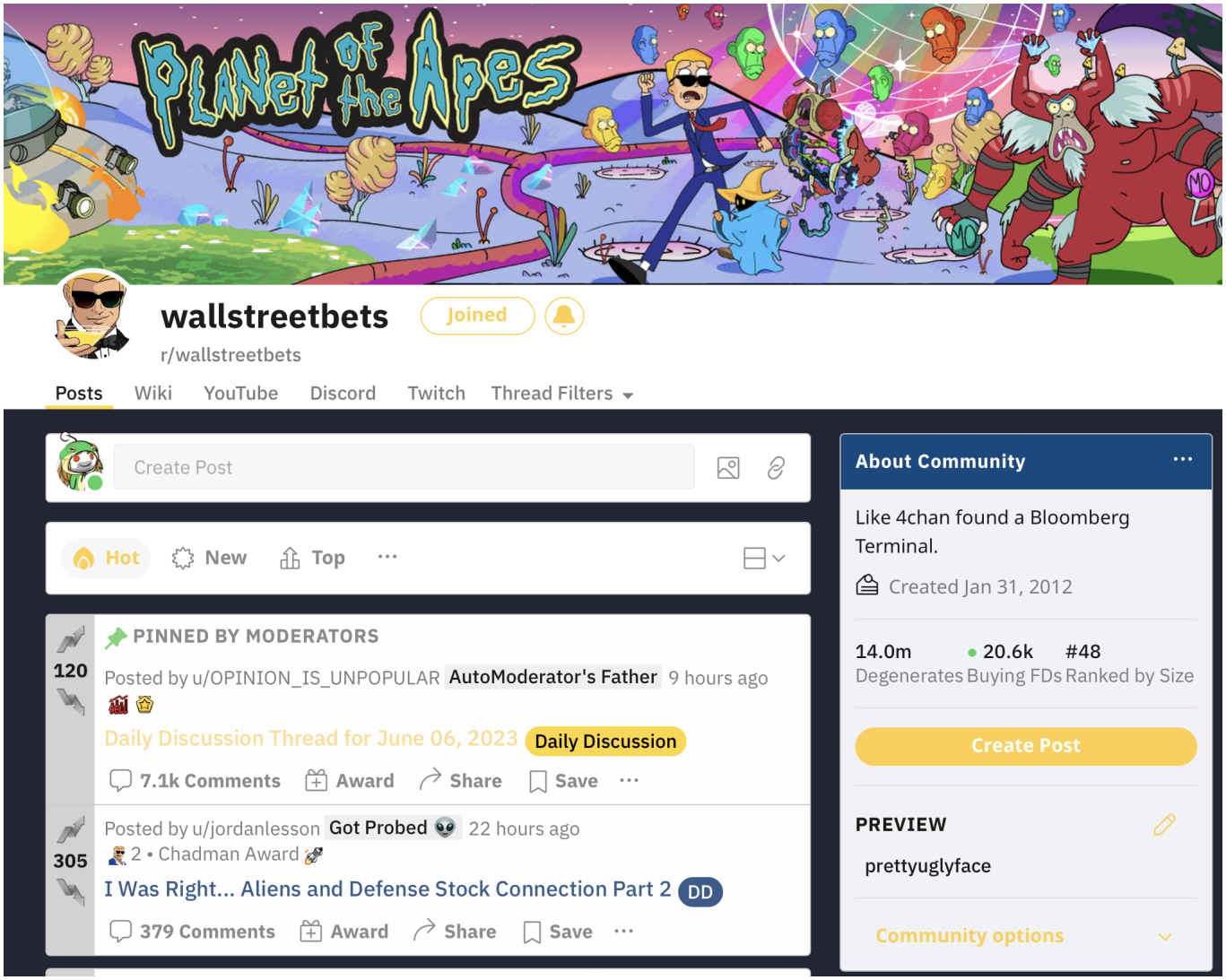


Figure A.1: WSB Screenshot

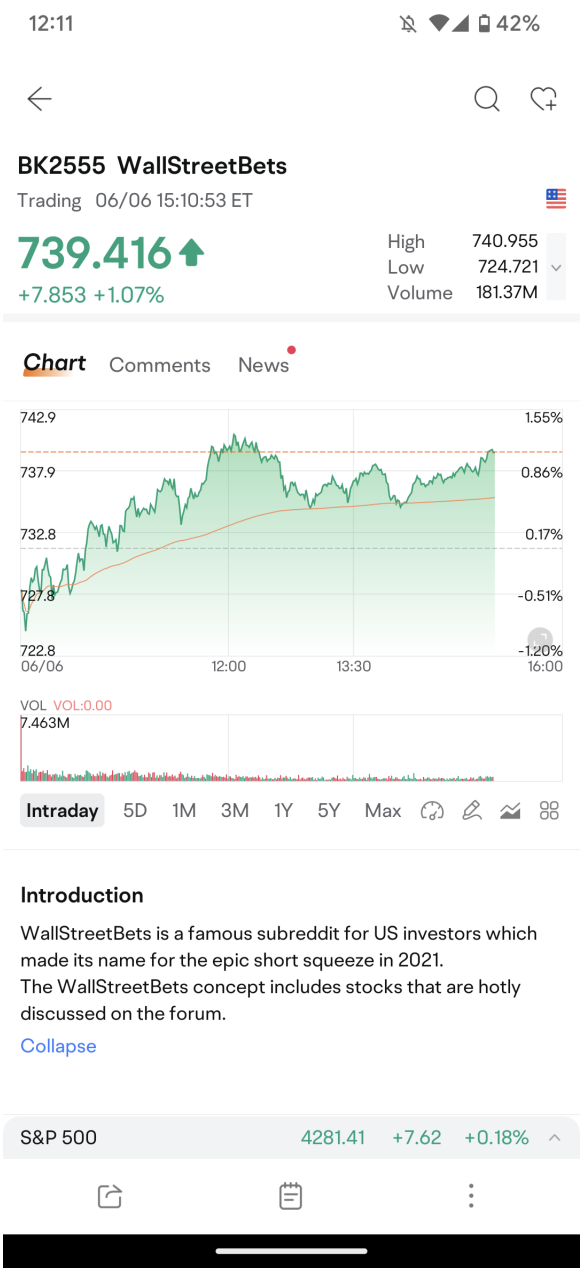


Figure A.2: Reddit WSB Index on trading platform MooMoo

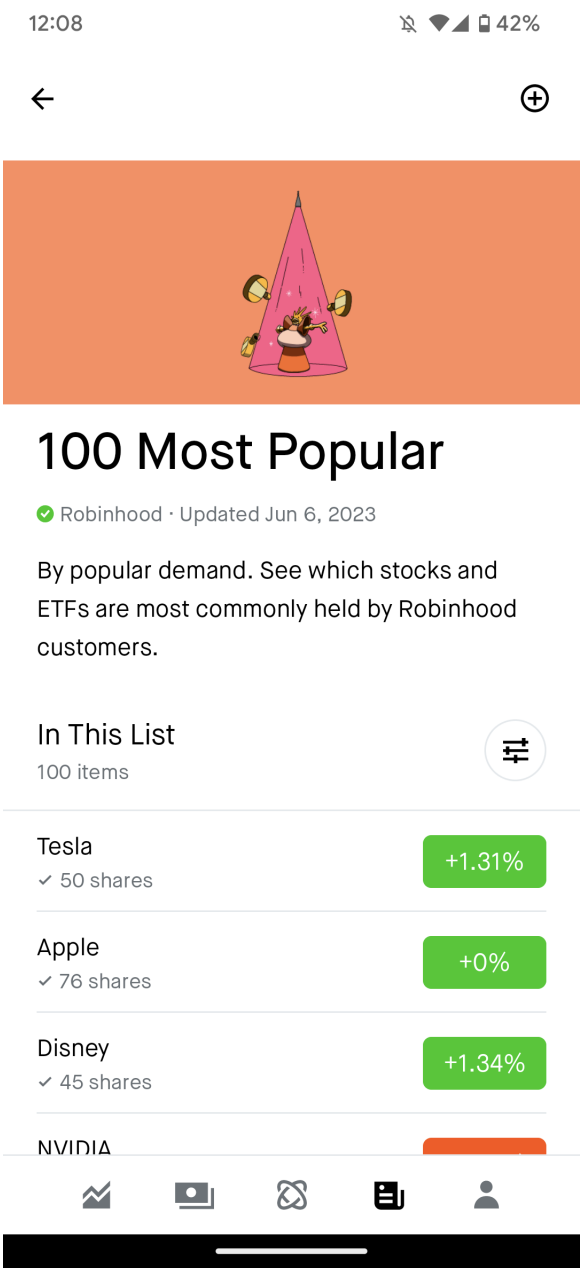


Figure A.3: RH 100 list on the RH mobile application