Essays in Corporate Finance: The Role of Social Media in Financial Markets

MUSTABSAR AWAIS
Registration No. 160130881

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Social Sciences
Management School

September 2021

Declaration

I, Mustabsar Awais, confirm that the Thesis is my own work. I am aware of the University's Guidance on the Use of Unfair Means (www.sheffield.ac.uk/ssid/unfair-means). This work has not been previously been presented for an award at this, or any other, university. However, parts of the thesis have been presented in various seminars and conferences.

Mustabsar Awais
Sep 2021

Acknowledgements

Firstly, I wish to express my deep and sincere gratitude to Dr. Junhong Yang, my research supervisor, for his overwhelming support and invaluable guidance. His commitment to achieving excellence in research and the quest for knowledge, is highly commendable. I am greatly indebted to him for this whole extraordinary learning experience and his innovative ideas, which helped shape every page of this thesis.

The support I received from the Management School must be acknowledged too, especially that of Dr. Jill Atkins, Dr. Martina McGuinness, and Dr. Jiao Ji for providing continuous support. Their guidelines and academic advice have enriched my PhD experience. I would also like to thank Mandy Robertson, who always responded to my emails and provided the appropriate guidance and administrative support during the PhD program.

To Prof. Dr. Anwar Ahmed, I owe a debt of gratitude. Without his unwavering support, this whole life-changing experience would never have been possible. Your guidance and teachings have always left lifelong and meaningful impressions, which will always be cherished with equal zeal and zest like today.

I am also thankful to Prof. Dr. Mohammad Nizamuddin for providing the opportunity to pursue this degree. My sincere thanks go to Prof. Dr. Rubina Tareen and Prof. Dr. Muhammad Shaukat Malik for offering continuous support during my PhD journey.

A special thanks to Dr. Ahmed Timsal for his extraordinary support and guidance, most noticeably for always standing beside me through thick and thin. I would also like to acknowledge the love and support from my younger brother Abdul Saboor Khan.

Finally, I would like to express my heartfelt gratitude for my wife who stood by me and assisted me throughout this PhD journey. Without her support, companionship, and unconditional love, this whole venture would have not been possible - Yes, we did it!

Abstract

In this thesis, I examine the role of investor-oriented social media platforms in the financial markets. This thesis comprises three standalone yet interconnected chapters in the field of corporate finance, which investigates and discusses three vital constructs of attention allocation in the financial markets.

Chapter two focuses on the attention allocation by investors on social media and its consequences for investors in the financial markets. Using more than 32 million tweets and a market microstructure dataset, this chapter investigates the impact of social media attention (SMA) on investors' trading behaviors. Through a battery of experiments, I find that SMA is a unique proxy of retail investors' attention and is different from other direct proxies. Aggregate data suggests that SMA can predict the financial markets, i.e., higher SMA results in short-term price pressures generated from more buys than sells. However, short-term price pressures are reversed the next day. Using social media heterogeneity, I test the information diffusion hypothesis. The empirical findings suggest that the impact of social media is amplified within the large social network. Moreover, tweets by verified users, replies, and retweets further increase the credibility of the information. This chapter contributes to the emerging body of knowledge that investigates the impact of social media attention on financial markets and provides valuable insights for a diverse set of market participants, mainly –retail investors.

In chapter three, I investigate whether disagreement on StockTwits provides firm-specific information. Using supervised machine learning approaches and a novel dataset, I predict investors' recommendations and measure disagreement among investors on StockTwits. The findings suggest that an increase in investors' disagreement results in a drop in return synchronicity. The negative impact of investors' disagreement on return synchronicity suggests higher inflows of firm-specific information. In line with this view, I find that disagreement improves price informativeness by increasing the price leads of earnings. Further empirical evidence suggests that the negative impact of disagreement on return synchronicity is more pronounced for firms with less transparent information environments and higher salience on StockTwits.

In chapter four, I examine the role of investor-oriented social media platforms to predict crash risk. Using the investor-level novel dataset from StockTwits and developing a unique proxy of investors' update in sentiment i.e., sentiment oscillations, I find that sentiment oscillations on StockTwits are significantly and positively related to the firm-level future crash

risk. These results remain consistent after using a battery of tests to account for unexpected market events, heterogeneity of investors, and endogeneity concerns using the instrumental variable approach. Further tests suggest that the impact of sentiment oscillations is more pronounced in firms with a lower level of accounting conservatism, less analyst coverage, less product market competition, and positive market sentiment. Overall, this study highlights the significance of social media platforms for investors and sheds light on the behavioral explanation of firm-level crash risk.

Table of Contents

List of Figures

List of Appendices

List of Tables

*C*hapter 1

1. Introduction

1.1. Background

The role and use of social media are evolving at lightning speed and have gradually transformed the communication channels, particularly during the last decade. It has promoted the concept of networked society and facilitated intra – and inter-organizational activities among peers, businesses, consumers, financial institutions, learners, and investors. Furthermore, it has significantly reduced the communication barriers among different cultures, demographics, and geographic locations by providing the opportunity to express and share viewpoints and opinions (Blankespoor, 2018; Bonsón & Flores, 2011; Ngai et al., 2015). Like any other segment of society, social media has an impact on the way investors consume information in the financial markets.

Early studies in accounting and finance established a weak link between social media postings and stock returns (Tumarkin & Whitelaw, 2001; Wysocki, 1998). However, Antweiler and Frank (2004) provided corroborative evidence that internet stock message boards and disagreement among these postings are positively associated with the trading volume. After the emergence of Facebook and Twitter in the late 2000s, the realm of social media has extended exponentially. This has also encouraged researchers and academicians to explore the emerging role of social media in financial markets, trading behaviors, etc. For example, Bollen et al. (2011) found that Twitter's mood predicts the stock market and, therefore, can act as a tool to measure market sentiment.

Similarly, Luo et al. (2013) examined the impact of using social media on firm equity value and found a weak association but strong just-in-time predictive power. However, using data from an investor-oriented social media platform, Chen et al. (2014) presented compelling evidence that investor-oriented social media platforms can help investors analyze the financial markets and access firm-specific information at low/no cost. More recently, Chen et al. (2019) found that CEOs' and CFOs' participation on social media platforms increases the flow of firm-specific information in the financial markets and facilitates executives to promote their firms.

It is, however, noteworthy that investors' sentiment and mood prediction models are not new. Saunders (1993) provides evidence that investors' sentiment is directly influenced by the weather conditions and predicts positive returns on sunny days. Using the same approach but using world stock exchange data, Hirshleifer and Shumway (2003) find that the investors' sentiment varies based on weather conditions. Similarly, Edmans et al. (2007) provide evidence that sports sentiment can influence the market returns of the host country. Furthermore, Hwang (2011) examines how country-specific sentiment impacts the investors' choice in the USA to invest in stocks from certain countries. However, from these previous studies, what distinguishes social media platforms is that it provides direct evidence to observe market- and firm-wide sentiment, investors' heterogeneity, and salience of information signals. Also, it provides a unique opportunity to observe continuous time series of the impact of the firm- and market-level events to influence investors' trading behaviors.

Some of the more recent studies have examined the role of social media platforms for investors from a behavioral perspective. For example, Azar and Lo (2016) find that tweets posted during and after the Federal Open Market Committee (*FOMC*) meetings can predict abnormal returns and guide investors to diversify their risk to earn higher portfolio returns. Cookson and Niessner (2019) provide evidence that within and across group disagreement among investors is associated with abnormal trading volume. Their study highlights investor-oriented social media platforms' significance and differential interpretation of information signals by investors using different investment approaches. Similarly, using data from Twitter, Nekrasov et al. (2021) concluded that an increase in the salience of tweets increases the tweets' audience. Therefore, to attract investors' attention during earnings announcements, firms are more inclined towards using visuals in their tweets to increase the salience of their information signals in the financial markets.

These prior studies provide substantial evidence to establish the growing role of social media in the financial markets. Recently, Hirshleifer (2020), in his presidential address at the American Finance Association, highlighted the role of social media and social interaction among investors and said:

"Social economics and finance recognizes that people observe each other and talk to each other, where talking includes written text and social media. A key but underexploited intellectual building block of social economics and finance is social transmission bias, the systematic directional modification of ideas or signals as they pass from person to person."

1.2. Research Motivation

Academic research and discussions on social finance have been going on for several years now. For example, Shiller (1992) argued that investing is a social activity where investors interact. In similar lines, Hong et al. (2004) also found consistent evidence that social interaction among investors increases participation in the stock market. Recently, Hirshleifer

(2019) provided a theoretical framework to examine the role of social interactions among investors in the financial markets.

Previous studies in this domain have mainly focused on the power of social media to predict volume and returns (Azar & Lo, 2016; Sprenger et al., 2014), firm-specific events (Mazboudi & Khalil, 2017; Nekrasov et al., 2021), consequences of divergence and convergence of investors' opinion with traditional media (Al-Nasseri & Menla, 2018; Giannini et al., 2019), disagreement among investors (Cookson & Niessner, 2019) and using as an information channel for firms to reduce information asymmetry in the financial markets (Blankespoor et al., 2014).

Motivated by these research findings, we aim to explore the following research questions. First, which segments of the financial market are influenced by social media? Second, does interaction among investors on social media platforms increase the flow of firm-specific information in the financial markets? Finally, to consolidate our findings, we test whether social media platforms for investors empower investors to anticipate firm-specific stock price crashes in the future.

1.2.1. Attention Allocation in the Financial Markets

In the extant literature that explores the role of social media in financial markets, some of the recent studies examining the role of attention allocation by investors provide valuable insights. For example, Hirshleifer and Teoh (2003) and Hirshleifer et al. (2011) highlight the significance of limited attention and argue that due to scarcity of cognitive resources (Kahneman, 1973), investors' in the financial markets need to allocate their attention efficiently. Similarly, Yuan (2015) provides evidence that market-wide attention events predict investors' trading behaviors in the financial markets. In the case of social media platforms, the attention allocation process is entirely different from market events. With the offering of Cashtags service by Twitter and the emergence of investor-oriented social media platforms, the challenge is locating firm-specific attention in a single post with multiple cashtags. Some of the prior studies have attempted to explain their social media data collection and cleaning process, however, this requires further details to unleash the true influence of social media in the financial markets. In this study, we aim to fill this gap i.e., following Evans et al. (2019) we clean Twitter dataset to predict investors trading behavior in the financial markets.

This study greatly benefits from the seminal work of Kahneman (1973), who argues that attention is the crucial ingredient to encode information signals in the financial markets. First, attention allocation is a reflex action by the market agents. However, due to the scarcity of cognitive resources, the attention allocation process must be as efficient as possible. Second, to process the information signals, market agents need to allocate further resources to interpret these information signals correctly. Thus, attention and effort are required to encode the information signals. In Figure 1, we provide visual guidance on this from Kahneman (1973).

[*Insert Figure 1.1 here*]

Due to the scarcity of cognitive resources and the instantaneous flow of information, the attention allocation to information signals that can increase investors' marginal utility in the financial markets is challenging. In this regard, Fiske and Taylor (1991) argue that the salience of information signals differentiates among information signals based on their perceived utility for investors in the financial markets. Therefore, the higher salience of information signals guides investors to allocate their attention efficiently. Motivated by these findings, the use of social media data emerged as a natural choice, as it facilitated us to directly observe the salience of information signals compared to any other dataset used for examining the role of attention allocation in the financial markets.

1.2.2. Consequences of Attention Allocation in the Financial Markets

There is a binary outcome after the investors allocate their attention to the financial markets. They can either agree with the information signals, thereby increasing their marginal utility to consume firm-specific information. Or they might disagree, thus increasing the marginal utility of other investors' information signals to update their analysis. In both cases, consumers and providers of information signals benefit each other. The extant literature in economics and finance argues that investors in the financial markets have different tastes and preferences. Consequently, such differences result in disagreement instead of agreement in the financial markets (Fama & French, 2007; Rubinstein, 1993). In a similar research line, Kandel and Pearson (1995) argue that differential interpretation of information signals by the investors in the financial markets increases the disagreement. As a result, investors update their economic models based on their learnings from such disagreements.

Overall, the disagreement literature in finance provides unified evidence that investors agree to disagree, and such disagreements among investors motivate the investors to actively participate in the financial markets (Banerjee & Kremer, 2010; Carlin et al., 2014; Cen et al., 2017; Hong & Stein, 2007). Although the existing literature highlights the significance of disagreement among investors, it does not provide any evidence of whether social interaction among investors results in a higher inflow of firm-specific information in the financial markets. This study aims to fill this gap.

1.3. Data Collection

In this doctoral thesis, I have relied upon multiple databases to complete the research analysis in the respective chapters. First, I used financial databases linked with Wharton Research Data Service (WRDS), Thomson Reuters Eikon, and Bloomberg available via Sheffield University Management School database services. Second, I created one of the largest social media databases on the virtual servers provided by the University of Sheffield IT services. Social media data is harvested from Twitter and StockTwits after approval from The University of Sheffield Research Ethics Committee via application no. 019102. Further details about these social media datasets are as follows.

1.3.1. Twitter

Twitter is one of the largest social media platforms in the world. It started its cashtags service in 2012. A Twitter cashtag is defined as a dollar sign with the company ticker ($AAPL). For data collection from Twitter, I have developed a customized programming application (CPA) that connects with the Twitter application programming interface (API) and downloads data based on pre-defined endpoints[1]. The CPA collects data using the following filters. 1) The tweet[2] must contain the firm's cashtag in our sample. 2) The tweet must contain only one cashtag. This is in line with Kahneman (1973) framework, to ensure that instead of shared attention the users are only discussing the firms with the specific cashtags. 3) The data collection period is from 2012 – 2016 (5years). To synchronize the Twitter data set with U.S capital markets opening and closing times, the timestamps of tweets are converted to Eastern Time (ET). A key difference between our social media datasets and the datasets used in previous studies is that I have collected every piece of public information related to the tweets. For example, users' details consist of users' public information, their followers, following, posts, etc. Tweets' details consist of the number of likes, retweets, and replies, etc. Moreover, since a tweet is a collection of words, hashtags, cashtags, mentions, URLs, and media links, it

[1] Twitter developer platform has evolved continusouly, further details about the Twitter's search endpoints can be found here: https://developer.twitter.com/en/docs/twitter-api/search-overview.
[2] According to Twitter API documents, it is not possible to collect all tweets since every time a tweet search gives tweets from a different bucket. Given that, we replicated our programming application several times. During the final run, we saw a few additions in our tweet database as compared to the first run. We did not collect any deleted tweets, or any information blocked by the owners of tweets.

is essential to parse[3] tweets to collect this information and save it in the database. Overall, with this process, I collect approximately 80 million tweets, posted by 1.2 million unique users discussing more than 5000 US-listed companies. I use the Twitter dataset in chapter 2 and explain further details about the additional filters I used to match the cashtags with the sample firms in the same chapter.

1.3.2. StockTwits

The second data source of the social media data in this study is the StockTwits.Unlike Twitter, StockTwits is an investor-oriented social media platform. StockTwits is by far the largest social media community for investors and traders, with more than 3 million members, 5 million monthly messages, and 3 million monthly visitors.[4] The main user interface of StockTwits is user-friendly; investors can post Twitter-like messages up to a 140-character limit.[5] One of the distinguishing features of StockTwits is investors' user profiles, where any investor can volunteer to disclose their asset choices, investment approaches, and investment term preferences. Moreover, investors can create a customized watchlist to view StockTwits' ideas directly relevant to their investment preferences.

To collect data from StockTwits, I have developed a customized programming application (CPA) to connect with StockTwits API and collect multiple data points based on numerous iterations. StockTwits relies on cashtags (for example, $AAPL) as company identifiers. I use the same filters as in Twitter for the data collection on StockTwits except for the data collection period, which is 2012 – 2018 (7 years). However, later on, I identified some technical issues in data collection in the year 2012 and 2018. Therefore, To this end, to ensure that data are relevant, I only keep data from January 2013 to December 2017. Overall, with this process, I collect approximately, 38 million ideas, posted by 297,000 unique users, discussing more than 8000 companies listed in various US stock exchanges. StockTwits data is used in chapters 3 and 4. Further details about using additional filters and data matching with the sample firms are explained in the respective chapters.

---

[3] For tweet parsing, we use python 3.6 and Natural Language Toolkit (NLTK) 3.4. Parsers can be cloned from our GitHub page https://github.com/Mustabsar

[4] https://about.stocktwits.com/

[5] On May 8, 2019, StockTwits increased their character limit to 1000 characters. However, this occurred outside our sample period.

1.4. Research Overview

In the second chapter, we examine the role of social media attention to predict investors' trading behaviors in the financial markets. Specifically, following the theoretical framework of Kahneman (1973), we establish a clear link between social media attention and net order flows of retail investors' trading. We accomplish this by using market microstructure data to differentiate between retail and institutional investors in the financial markets (Barber et al., 2008; Kumar & Lee, 2006; Lee & Ready, 1991). To further extend our analysis, we examine the interaction between the salience of information signals and social media attention. Consistent with the prior studies, we find that the salience of information signals pronounces the impact of social media attention. In addition to this, we highlight some of the critical challenges for researchers while using data from a large social network, i.e., Twitter. Overall, this chapter provides valuable insights for the researchers examining the role of large social networks in the financial markets.

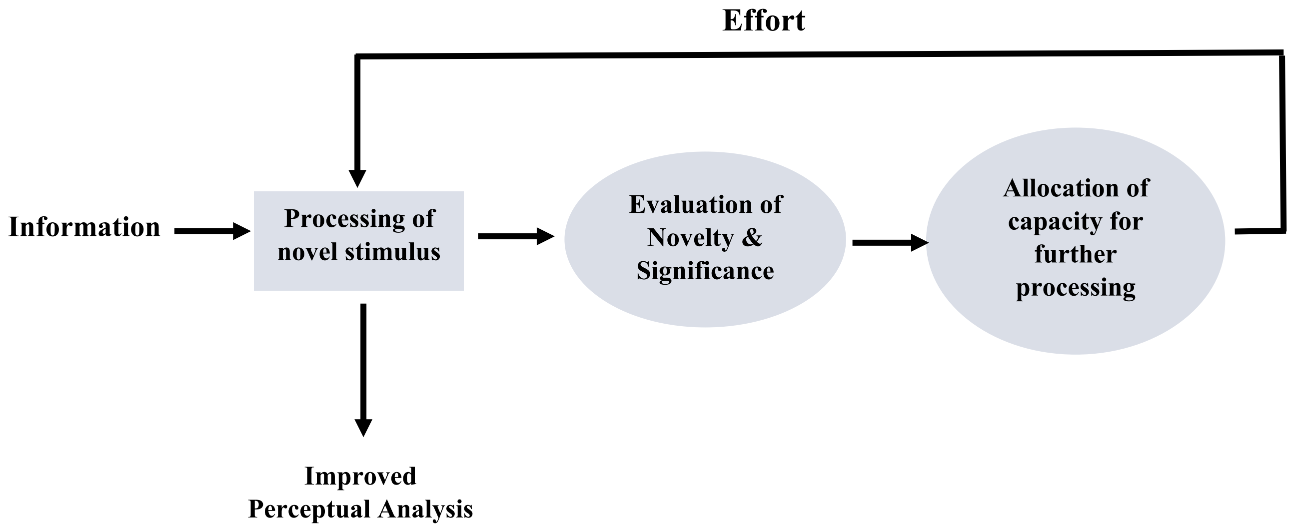
It is noteworthy that social media platforms offer investors the opportunity to interact and share their opinions and recommendations. However, the existing literature lacks evidence to explain if such interactions on social media platforms provide any value-relevant information to the investors in the financial markets. Therefore, in the third chapter, we extend our analysis and complement the existing literature by providing direct evidence that discussions on social media platforms increase the flow of firm-specific information in the financial markets. In this chapter, to extend the credibility of our findings, we rely on one of the largest investor-oriented social media platforms, i.e., StockTwits. Following the theoretical framework of Hong and Stein (2007), we find that overall disagreement among investors on social media platforms provides a unique opportunity for investors to consume value-relevant information and share their analysis.

Based on our findings in the previous chapters, we are keen to examine investor-oriented social media platforms' role to facilitate investors to predict firm-specific events in the fourth chapter. Following Hong and Stein (2003) theory of market crashes and recommendations from their companion paper (Chen et al., 2001), we examine whether the update in investors' sentiment on StockTwits can predict firm-level future crashes. Our research findings are consistent with the prior literature on crash risk and provide new evidence that social interaction among investors empowers investors to anticipate firm-specific events.

1.5. Thesis Structure

The thesis comprises of three stand-alone but interconnected chapters. In the second chapter, we examine the influence of investors' attention allocation in the financial markets. In the third chapter, we test the consequences of attention allocation by examining the association between disagreement among investors and the stock return synchronicity. The fourth chapter consolidates our findings from the previous chapters by examining investor-oriented social media platforms' role in predicting firm-specific future events. Chapter five concludes the thesis by providing overall conclusions and practical implications of the study.

Figure 1.1: Processing of novel stimulus (Kahneman, 1973)



Notes: The figure presents the two types of inputs to process novel stimuli, i.e., information and effort.

Chapter 2

2. Does Social Media Attention Affect Investors' Trading Behaviors?

2.1. Introduction

Given the massive publicity and coverage of social media, it seems that we are witnessing an entirely new era of information sharing. Social media, today, are seen as being responsible for a revolution,[6] creating panic in financial markets,[7] fueling political campaigns,[8] and at times punishing firms.[9] Within financial markets, their dynamic role has emerged as a centerpiece because their continuous stream of information offers a unique opportunity for market participants to sift through an abundance of information. Financial markets are driven by information and, by using cashtags, *Twitter* provides investors with an opportunity to share their opinions and collect information about specific stocks. Moreover, the breadth of information an investor can get on *Twitter* using hashtags and mentions is inimitable.

Social media have recently developed the concept of a networked society, broadly increasing the pace and variety of collaboration. The last few years have seen an emerging strand of research on social media and its impact on financial markets. For example, Bollen et al. (2011), using machine learning techniques, classified tweets and concluded that Twitter moods predict stock market returns; Blankespoor et al. (2014) suggested that tweets from company handles can help reduce information asymmetry; Sprenger et al. (2014) used 250,000 tweets to understand the association between Twitter sentiment and information aggregation on volume, returns, and volatility; and Azar and Lo (2016) used Federal Open Market Committee (FOMC) meetings' news on Twitter as an attention shock for financial markets and developed buy and hold portfolios.

Overall, previous studies have mainly focused on Twitter sentiment analysis. However, two key arguments may help finance scholars understand social media's role in financial markets and warrant further study. First, attention and sentiment are two different reactions to

[6] Refer to Howard et al. (2011) for further reading.
[7] A tweet posted by hacked Associated Press Twitter handle (@AP) on 23 April 2013 created havoc on Wall street and as a result the S&P 500 index was reported to have lost $136.5 bn.
[8] Donald Trump, after winning the US presidential election, tweeted on 12 December 2016 about the F-35 fighter jet program of Lockheed Martin; Lockheed Martin, one of the largest defence sector companies, lost $4 bn in value due to a share price fall. A single tweet cost $28.4 m. per character to shareholders of the company.
[9] In a viral video posted on social media, a passenger was dragged off a United Airlines flight, the footage prompted outrage globally and cost United Airlines approximately $1 bn. It not only prompted great debate on social media but also US lawmakers to initiate another passenger safety bill in the airline industry.

stimuli. For example, greater attention does not necessarily mean positive sentiment and vice versa. Second, Twitter is a large social network with several features distinguishing it from other popular social media platforms for investors. These features may increase the salience of information signals via tweets. However, a potential challenge for researchers is to address noise in tweets without losing the salient features of tweets and a large amount of data. Therefore, the focus of this study is on specifically addressing these two key issues by explaining the role of social media attention (SMA) in financial markets and the role of social media heterogeneity to increase the salience of information signals.

The efficient-market hypothesis states that information is already incorporated into stock prices when publicly available, i.e., investors pay significant attention to all information available in the market. This argument by Fama (1991) contradicts the attention framework by Kahneman (1973), who states that attention is a scarce cognitive resource and a fundamental ingredient for investors to respond to any stimuli before making any decisions. Therefore, given that not all investors have equal access to information sources, limited attention may impact different stocks. While making investment decisions, a rational investor can overreact or underreact to available information, affecting stock prices differently than stated in the efficient market hypothesis (De Bondt & Thaler, 1985). Similarly, considering the variations in social media platforms and the quality of information shared by them, attention allocation by investors depends on several other social media characteristics, such as an extensive social network (Han & Yang, 2013). Thus, the impact of limited attention can vary greatly from one social media platform to another depending on geographic location (Baik et al., 2016) and users' influence (Bhagwat & Burch, 2016) on a social network.

This study greatly benefits from the seminal work of Kahneman (1973) and the recently much discussed *wisdom of crowds* hypothesis by Surowiecki (2005). The study endeavors to address an existing research gap by offering substantive evidence on how large crowds allocate their attention in financial markets. Therefore, using the wisdom of the Twitter crowd, we investigate the impact of social media attention (SMA) on investors' trading behaviors. Using Twitter as a data source, our data set meets all four characteristics of large crowds (Surowiecki, 2005). Using Twitter cashtags as a direct proxy for investors' attention, our data set fulfills the criteria set out by the attention framework (Kahneman, 1973).

Our study is based on one of the largest Twitter data sets, containing more than 32 million tweets posted by more than 1.1 million users. Using natural language processing (NLP) techniques, our Twitter sample offers a discrete measure of SMA for every stock-day level of information. The Twitter data set used in this study is one of the most extensive collections of tweets with cashtags that cover 2,675 U.S listed firms from 2012–2016. The breadth of our Twitter data set allows us to investigate further the impact of social media heterogeneity. Unlike relying on traditional sentiment approaches, this study relies on the actual content of tweets and parses them based on their various characteristics, such as media content, URLs, and mentions that appear in the text of tweets.

A prominent feature of Twitter is that it offers real-time access to information. To account for tweets' real-time impact, we divide the Twitter sample into trading and non-trading hours. Unlike relying on volume to understand trading behaviors, we classify trades into buy/sell using the DTAQ data set and compute net order flow (NOF) as a proxy for investors' trading behaviors. *NOF* has been used in previous studies[10] to infer retail investors' sentiment, the cost of information asymmetry, and market-wide attention. Earlier studies by Da et al. (2011) and Ben-Rephael et al. (2017) recommend using Google Search Volume Index (GSVI) and Institutional Attention via Bloomberg terminals as direct proxies for retail and institutional investors, respectively. Therefore, due to the nature and quality of information shared on Twitter, it is equally important to understand which segment of the financial market is represented by large social networks such as Twitter. Using market microstructure data, we group trades into small, medium, and large; where small trades represent retail investors, and large trades represent institutional investors (Yuan, 2015).

Our research findings suggest that SMA influences retail investors' buying behavior. Consequently, it results in temporary price pressure, which is reversed the next day along with buying behaviors. Similarly, institutional investors' trading behaviors are not affected by SMA since institutional investors already have access to reliable resources as compared to retail investors. These findings are consistent with the notion that retail investors have a choice to select the best stock at the time of buying. However, they can only sell what they own (Barber & Odean, 2008). These findings support the evidence that not all investors can process the information on merit and, as a result, retail investors choose the stocks with the highest level of SMA at a given point in time. Using alternative proxies for retail investors, the findings

[10] See, for example, Chordia et al. (2017) and Chelley-Steeley et al. (2019).

remain consistent and, as a result, one standard deviation increase in SMA results in a 7.1% increase in NOF in small trades groups.

Twitter is used by investors, and it is popular among various media outlets to release breaking news and among firms to share firm-related information. Therefore, we need to understand the impact of SMA in the presence of significant news developments such as flash stories and traditional news stories about U.S. listed firms. Our findings suggest that SMA plays an important role when various media outlets release flash news as compared to traditional news stories. These findings offer evidence that social media increase the diffusion of just-in-time information during trading hours, resulting in attention-induced price pressures. Given our Twitter sample's breadth, we test whether the tweets contain media content along with cashtags. Our substantive findings offer evidence that tweets with embedded media content have a significantly larger impact on NOF, which remains consistent even in the presence of traditional media coverage. To understand the effect of attention-induced price pressures, we use Fama and MacBeth (1973) regression. Our findings remain consistent, in that SMA results in attention-induced price pressures, with one standard deviation increase in SMA resulting in a 9.80 bps increase in stock returns. Our findings remain consistent for small firms, firms with less analyst coverage, and firms with higher retail ownership, which further suggests that price pressures are significant in small firms in which retail investors mostly trade. These findings complement our previous finding that the impact of SMA on NOF results in attention-induced buying behaviors.

To examine the impact of the salience of firm-level tweets, we divide our Twitter sample based on distinguishing features, such as tweet-specific and user-specific characteristics. In the case of user-specific characteristics, we define information diffusion as the number of users a tweet can reach (*Tweet Reach*) in the Twitter network and the number of distinct users (*Users Concentration*) who are actually posting tweets with cashtags during trading hours. We test our hypothesis on both of these scenarios. To understand the impact of tweet reach, we divide our sample into terciles based on the lowest to highest tweet reach. Our results present a threefold incremental impact of SMA on NOF, suggesting that if the information is diffused within a large social network, the overall impact of information increases along with SMA. Tweet-specific characteristics are divided into three groups based on replies (*Discussion*), retweets, and user's account status (*verified/ non-verified*). Our results show that tweets with large discussion threads have a significantly larger SMA impact for retail investors compared to no/low discussion threads. Furthermore, these results show that users with a verified account status have a substantially larger SMA impact than users with a non-verified account status, thus suggesting that information credibility on social media plays an important role.

These findings are robust for a wide range of control variables and alternative measures of attention proxies. In a recent study by Da et al. (2011), they presented Google Abnormal Search Volume Index (ASVI) as a direct proxy for the attention of retail investors. We test the impact of ASVI in our research settings and find that ASVI is a direct proxy of retail investors (as in Da et al. (2011)). However, SMA captures attention in an entirely different fashion, i.e., real-time attention with the salience of information signals. We use Lee and Ready (1991) trade classification method to classify trades as buy/sell. However, our results remain consistent with different trade classification methods by Ellis et al. (2000) and Chakrabarty et al. (2007). To further differentiate between retail and institutional investors, we divide trades into three groups, i.e., Small, Medium, and Large trades[11]. Our summary statistics for different trade groups are consistent with previous studies (Barber et al., 2008). To validate our findings, we also run our model for SMA during non-trading hours. The results remain consistent but with weak coefficients, further suggesting that SMA is more of a real-time attention proxy. Lastly, we run our regression model during trading hours and include earning announcements as attention shocks as another variable of interest. Our results remain consistent even in the presence of attention shocks.

This study contributes to the emerging literature on the role of social media in financial markets. Our findings should be of broader interest to scholars in the field of market microstructure and behavioral finance. Unlike previous studies, which mainly focus on the sentiment analysis of social media text, this study complements existing research on social media. It highlights the importance of user-generated aggregated social media content and its impact on investors' trading behaviors by presenting a unique proxy for attention. Most importantly, with a recent trend towards investing in social media platforms, our study supports the *wisdom of the crowd* hypothesis (Surowiecki, 2005) by drawing a sample from an extensive social network, i.e., Twitter.

[11] Trades classifications are discussed in section 2.4.

This study extends existing research on social media (Antweiler & Frank, 2004; Tumarkin & Whitelaw, 2001) by offering substantive evidence of social media's impact on retail investors' trading behaviors. Another contribution of this study is to the market microstructure literature by studying it along with social media for a large sample of US listed firms. The study benefits from the use of transactional data such as DTAQ to classify trades. This is the first study to use net order flow on a comprehensive cross-sectional sample of a five-year period from NYSE/AMEX/NASDAQ to understand the impact of SMA on investors' trading behaviors. Unlike existing research, which uses volume and stock returns as dependent variables, the natural appeal of using net order flow in this study stems from the fact that SMA is used as a direct proxy for trading behaviors and market sentiment (buy/sell).

The rest of the paper is divided as follows: Section 2 reviews the existing literature and explain the hypothesis development; section 3 describes the sample selection and research design; section 4 presents our empirical analysis; section 5 presents robustness checks and external validity tests, and section 6 concludes the paper.

2.2. Literature Review and Hypothesis Development

The ability to access information and interact with others on different social media platforms has amplified the impact of information on our lives. In recent years, it has proved to be an influential medium of information sharing, as it has the power to influence public opinion. This was evident during the recent US presidential election (Bovet & Makse, 2019) and Brexit (Grčar et al., 2017) in the UK. Due to their nature, social media can diffuse information faster than traditional mediums (Foroozani & Ebrahimi, 2019; Kümpel et al., 2015; Liang, 2018). However, while there is a consensus that social media have dramatically reduced information barriers, there are also growing concerns that social media have given rise to misinformation and fake news, which can affect public opinion in an unanticipated fashion.

With the recent upsurge in social media platforms and the numbers of users of them, it is not difficult to see that they have gradually become crucial for investors. Some earlier studies discussed the concept of 'Internet stock message boards' and their impact on financial markets. For example, Tumarkin and Whitelaw (2001) used *raggingbull* and argued that Internet stock message boards do not predict industry-adjusted returns even after using text analysis to classify messages as buy/sell signals. However, in their study, Antweiler and Frank (2004) presented a compelling argument that although Internet stock message boards predict stock returns, their economic impact is negligible. However, a potential constraint of these studies is that the social media platforms discussed were not very common amongst investors at the time these studies were conducted, therefore raising concerns about their respective data set(s) and validity of their results.

Within the last decade or so, considering the nature and popularity of these platforms, Twitter has become one of the most popular social media platforms for investors. Being a leading social media and networking platform, Twitter offers a unique opportunity for investors to interact with each other using Twitter cashtags (Evans et al., 2019). Some studies have focused on explaining how Twitter affects financial markets using Twitter moods (Bollen et al., 2011), Tweet sentiment (Sprenger et al., 2014), and information aggregation (Azar & Lo, 2016). Blankespoor et al. (2014), using a Twitter data set, argued that information shared by companies using their Twitter handles could help to reduce information asymmetry in financial markets. In their study, Sprenger *et al.* (2014) associated Twitter activity with abnormal returns and market volumes. More recently, Mazboudi and Khalil (2017) found that social media play a vital role during major corporate investment decisions such as mergers and acquisitions. They further suggest that social media can be used to enhance stock price stability. Although these studies offer unique insights into social media's impact on financial markets, they do not explain how investors consume information in a large social network. Social media do provide an abundance of information; however, not all information attracts the attention of investors, as attention is a limited resource and can only be generated with some effort. In this regard, the aggregate impact of attention becomes more important to explore than individual attention.

2.2.1. Limited Attention and Social Media

Kahneman (1973) argues that there is always a mechanism that explains the significance of any stimulus. Since we choose to decide which stimuli we are paying attention to, in return, we allow those stimuli to control our behaviors. However, how much attention is devoted to any stimulus will depend on its novelty and complexity. More importantly, the individual also needs to determine their preferences to allocate their attention. Therefore, attention is a scarce cognitive resource. Complementing the attention theory from Kahneman (1973) with existing literature on social media offers unique insights into how humans allocate their attention and explain how social media play an essential role in our lives.

One of the main features of Twitter is the ability of users to express their views and opinions via tweets. However, these tweets can be used as an opportunity to build a social media image and attract more users to a profile, who are later potential followers of such profiles. Zhao and Rosson (2009) argue that Twitter offers real-time access to information and an opportunity to interact with others by sharing different tweets informally. However, users with more followers may significantly influence those with fewer followers (Cha et al., 2010). Therefore, users allocate more attention to those with more influence on social media, and this mechanism helps them sift through relevant information.

By reviewing some recently published research studies in this domain, we have identified some critical research gaps that need to be addressed. For example, in some of these studies, researchers only focus on how Twitter can predict financial markets (Becker & Nobre, 2013; Blankespoor et al., 2014; Bollen et al., 2011; Oliveira et al., 2017; Yu et al., 2013). With regard to attention, Da et al. (2011) used Google Search Volume Index (GSVI) to predict financial markets and explain how GSVI can be used as a direct proxy for investors searching for firm-related information on Google. Similarly, Chen et al. (2014) used stock tickers as an explicit proxy for investors' attention and examine social media's effect on financial markets. Therefore, most of these studies are unable to explain how users allocate attention.

Similarly, there are issues related to sample sizes and data validation as well. For example, Evans et al. (2019) suggest that Twitter cashtags and stock tickers on various social media websites can confuse firms registered on multiple stock exchanges. As a result, two different companies can have similar stock tickers in two different stock exchanges. Therefore, one needs to use additional scientific methods to make corrections to their sample. Some other studies (Giannini et al., 2019; Ranco et al., 2015; Sprenger et al., 2014) that use Twitter cashtags also suffer from cashtag collision bias.

Furthermore, in the context of attention theory (Kahneman, 1973) and how individuals make decisions (Tversky & Kahneman, 1974, 1981) to allocate their attention, cashtag collision can be an obstacle. Similarly, in the case of Twitter sentiment analysis, cashtag collision bias arises when a tweet with multiple cashtags cannot convey the sentiment about a single company. This study has attempted to address these research gaps and issues, as it complements the existing literature, caters for data validation issues, and ensures that the sample does not suffer from collision bias.

2.2.2. *Retail and Institutional Attention*

Information needs to be unique to attract investors' attention (Kahneman, 1973). Considering the nature of financial markets, there are various types of investors who trade in them. To investigate the impact of a specific attention proxy, we must have a direct and unique proxy for investors' attention. Recent studies such as Da et al. (2011) argue that GSVI is a unique proxy for retail investors. Similarly, Engelberg et al. (2012), in their study, suggest that the tips offered by *Jim Cramer,* in his popular TV show *Mad Money,* affect individual investors' trading behaviors. Ben-Rephael *et al.* (2017) show that as Bloomberg terminals are widely used by institutional investors, the news consumption indicators on these terminals offer a unique proxy for institutional investors' attention. Overall, there is a consensus in the literature that attention proxies can be differentiated based on accessibility and effectiveness. Since institutional investors have more resources and expertise to process information as soon as it arrives in the market, they take their lead from retail investors and play a pivotal role in permanent price adjustments within financial markets (Ben-Rephael et al., 2017).

Social media offer a unique opportunity for both institutional and retail investors in financial markets. Chen et al. (2014) argue that retail investors widely use social media commentaries to understand market behaviors and then predict financial markets. Similarly, Blankespoor et al. (2014) also state that firms use Twitter to disseminate information to investors quickly. In their study, Bhagwat and Burch (2016) found that firms use Twitter to influence investors so that they can improve market efficiency and eliminate any information friction in financial markets. Chen et al. (2019), using Tweets posted by CEOs and CFOs, suggest that social media reduce information asymmetry in financial markets and attract retail investors. Because of the nature of the information available in such social media streams, there is a consensus within relevant literature that such platforms can help retail investors (instead of institutional investors) gather information and mitigate information friction in financial markets.

2.2.3. *Wisdom of Crowds*

Recent years have shown an interest amongst researchers to investigate if crowd wisdom can predict financial markets. However, there is still a need to understand how crowds differ in their judgments as compared to experts. A recent study by Mollick and Nanda (2015), who used data from one of the largest crowdfunding sites, argues that crowds have the ability to make rational decisions. In several instances, crowds can outperform experts. Similarly, in their study, Azar and Lo (2016) argue that platforms with a large social network such as Twitter represent the (wisdom of) crowds and can predict financial markets and outperform other benchmark asset allocation strategies. Twitter is the largest social network for investors from diverse backgrounds and geographic locations (Baik et al., 2016).

Surowiecki (2005) suggests that to unleash the wisdom of crowds, crowds must meet four conditions: (1) The crowd must have a diversity of opinions such that everyone in the crowd has some unique information even though it may be an eccentric interpretation of known facts; (2) people' opinions are not based on their peers' opinions; (3) there may be some experts and amateur group members; and (4) an information aggregation mechanism exists for further understanding of crowd activity. Although there is no limit on the size of a crowd, the general understanding is that the bigger the crowd, the better it is, since a bigger crowd has more chances of meeting the conditions of a wise crowd, as compared to smaller crowds.

2.2.4. Market Microstructure and Social Media

Market microstructure data offer a unique opportunity for researchers to understand market trends and order flow in a more informed manner. As compared to any other proprietary data sets, the trade and quote (TAQ) data set offers full market coverage with a timestamp on trades. Barber et al. (2008) used a Trade and Quote (TAQ) data set from 1983–2000, segregated microstructure data into signed trades using Lee and Ready (1991) method, and found that retail trades could predict future returns. Yuan (2015) used the daily trade and quote dataset (DTAQ) to generate aggregate order flow and classify signed trades into small, medium, and large. He argues that large trades are a proxy for institutional investors' trading behaviors, while small trades are a proxy for those of retail investors. Antweiler and Frank (2004), while investigating the impact of Internet stock message boards, used a TAQ data set to generate trade value categories and 15-minute spread during trading hours.

One of the benefits of using market microstructure data is that we can calculate order flows to understand the presence of informed traders in financial markets (Holden & Jacobsen, 2014). Recently, much scholarly research has been devoted to investigating the impact of media coverage and stock returns, where researchers have used traditional proxies for attention such as abnormal turnover and returns (Barber & Odean, 2008; Fang & Peress, 2009). A higher volume may be caused by many buyer or seller-initiated trades, which have an entirely different impact on the market compared to total market volume (Yuan, 2015). Therefore, one needs to analyse buyer- and seller-initiated trades to understand the reasons for abnormal volumes and price pressures. In this regard, order flow can offer better insights into traders to differentiate between buy/sell signals in markets (Chordia & Subrahmanyam, 2004).

One of the most important pieces of information that can be ascertained from market microstructure data is the prediction of trade directions. In this regard, Lee and Ready (1991) trade classification method are one of the techniques most commonly used by researchers. Ellis et al. (2000) argue that the Lee and Ready (1991) method has an accuracy level of 81.05%, as compared to the quote rule and tick rule. However, they highlight a bias in large trade classifications when trades are executed inside quotes and suggest an improved algorithm for NASDAQ trades. Chakrabarty et al. (2007) suggested an alternative method for trades executed inside quotes and argue that the estimates of Lee and Ready (1991) and Ellis et al. (2000) are biased. Easley et al. (2016), using NASDAQ ITCH data, argue that the tick rule and bulk volume classification are relatively better classifiers for NASDAQ market data. However, Chakrabarty et al. (2015) also compared BVC, tick rule, and the Lee-Ready algorithm for a large sample of equities and argue that tick rule and Lee-Ready methods are the most accurate trade classifiers using NASDAQ data. In another study, Chakrabarty et al. (2012) tested traditional trade classifiers on short sales and argue that the Lee-Ready method is one of the most accurate trade classifiers, even for short sales.

At this point, it is also pertinent to mention that Barber et al. (2008) and Hvidkjaer (2008) found that trade classifications from the TAQ data set became less accurate due to decimalization in 2001. However, the question remains: can individual investors change their trading patterns because of decimalization? Yuan (2015) suggests that such a structural change is not possible in individual investors' trading patterns. Similarly, a recent study by Boehmer et al. (2017) suggests a new method to track the retail investors by choosing the right exchange code. It is pertinent to mention that we already use the similar exchange code in addition to other exchange codes as well.

2.3. Hypotheses Development

Limited attention is the ultimate consequence of an abundance of information in financial markets. Since time and resources are cognitive, it is impossible to allocate attention to every stimulus. Therefore, attention must be selective and demands effort as a primary source to process any stimulus (Kahneman, 1973). However, selective attention may result in narrow framing, which can further affect individuals' decision-making ability (Hirshleifer, 2001). As a result, individuals analyze information in an isolated manner, thus affecting their choices (Tversky & Kahneman, 1981). The framing effect has dire consequences for retail investors compared to institutional investors since they face stock selection problems while choosing from thousands of stocks. Some recent studies, such as those by Barber and Odean (2008) and Yuan (2015), provide further evidence to suggest that limited attention can affect investors' buying behaviors and consequently stock prices.

The general premise of this study is that retail investors are net buyers of attention-grabbing stocks. The rationale behind this premise is that investors usually have limited attention and resources. When they get the option to make a buying decision from among thousands of available stocks in the market, the task becomes challenging. Hence, limited attention and resources limit their ability to rank or analyze every stock (Barber & Odean, 2008). A recent study by Da et al. (2011) supports this argument. During the stock selection process, investors' decisions are based on the salience of the information signals they receive from various available resources. For example, prominent stimuli can attract more attention than complex ones that do not coincide with the conscious thoughts of individuals (Tversky & Kahneman, 1973). Salience plays a pivotal role in stock selection, and it has robust and wide-ranging effects (Fiske & Taylor, 1991). Moreover, the perception of every stimulus depends on its salience, whether it can attract more attention or not. Interestingly, retail investors do not face a stock selection problem when selling since they can only sell what they own. Overall, attention is the principal element to encode environmental stimuli such as financial information disclosures on social media (Blankespoor et al., 2014) or any other source of information disclosure (Hirshleifer & Teoh, 2003).

In the case of institutional investors, they have enough resources to allocate their attention and make informed decisions during the investment process. However, institutional investors face stock selection problems when selling (Barber & Odean, 2008). For example, hedge funds are actively involved in short selling. For those institutional investors who do not short sell, they may still suffer through stock selection problems when selling because they already own thousands of stocks as compared to retail investors. Moreover, the ability of institutional investors to analyze financial markets and make investment decisions outweighs the ability of retail investors. They can predict market events (Hendershott et al., 2015) because they have access to sophisticated resources such as Bloomberg terminals (Ben-Rephael et al., 2017), which are rarely available to retail investors.

To reassure themselves in the stock selection process, retail investors rely on various information resources that can facilitate them to sift through an abundance of information and make investment decisions. Chen et al. (2014) suggest that retail investors have started consuming information from social media platforms specifically meant for investors. Similarly, Cookson and Niessner (2019), in their study, concluded that investors' disagreements on social media platforms can predict market volumes and abnormal returns. Unlike any other traditional news services, such social media platforms give users the ability to read investment recommendations and analyses from investment gurus. Nisbett and Ross (1980) argue that individuals tend to prefer simple information and undervalue statistical information. This further supports the argument that individuals do not allocate their attention based on the economic significance of information (Hirshleifer & Teoh, 2003). Therefore, the limited information processing capacity of retail investors convinces them to consume information from such social media platforms, rather than processing it themselves for further analysis (Payne et al., 1993).

Our first hypothesis is based on the association between social media and retail investors. Retail investors suffer from stock selection problems. Social media platforms such as Twitter offer a unique opportunity to sift through an abundance of information via the use of Twitter cashtags. Therefore, information shared on Twitter has obvious implications for retail investors. Unlike previous studies, by using net order flow (NOF), our hypothesis is directly linked to the trading (buy/sell) behaviors of retail investors.

H1a: Ceteris paribus, social media attention (SMA) is positively associated with retail investors' trading.

Under different circumstances, individuals may under/overreact to various information sets available in financial markets (Payne et al., 1992), thus continuously updating their investment decisions. If SMA positively impacts retail investors' trading behaviors, the natural question to ask is whether SMA can affect prices. Hirshleifer and Teoh (2003) argue that financial markets react immediately to news events. Their argument also aligns with several short window event studies. However, recent technological developments and the instantaneous flow of information in financial markets make this phenomenon more interesting. The important question is how information is released to financial markets. Since social media are the quickest route for transmitting any information, this may have wide-ranging implications for financial markets. Since SMA is different from traditional media sources as it offers a real-time flow of information in the market accompanied by the salience of information signals, it should exacerbate the contemporaneous effect of SMA as compared to time sequencing's impact on the market returns. Therefore, to further investigate the impact of SMA on price pressures, we propose the following hypothesis:

H1b: Social media attention (SMA) is positively associated with the price pressures in the financial markets.

A distinctive feature of social media is their high market concentration and breadth of information. Social media platforms offer investors the opportunity to follow market sentiment and share any relevant information (Chen et al., 2014). Shiller (1995) argues that conversation plays an essential role in informational cascades, and social media disagreements can result in lengthy debates. Harris and Raviv (1993) found that disagreements in messages posted on various platforms by investors result in increased trading volumes. Considering the level of social media heterogeneity and the impact of information diffusion via social media, we develop our *Diffusion Hypothesis*. In their study, Dyck et al. (2008) suggest that two main characteristics of media are Diffusion and Credibility, which play an essential role in harming or harnessing a firm's reputational capital. Therefore, sometimes, firms try to assuage media demands by issuing updates more frequently, such as press releases (Ahern & Sosyura, 2014) and writing Wall Street Journal (WSJ) commentaries (Pound & Zeckhauser, 1990). In 2012, the U.S. Securities and Exchange Commission (SEC) said social media are landscape-shifting; therefore, investment managers are encouraged to use social media to disseminate any information related to their firms. In a recent study, Blankespoor et al. (2014) concluded that information dissemination via social media improves firms' visibility in financial markets and consequently is associated with higher liquidity. Sprenger et al. (2014) provide reasoning for this argument by stating that Twitter users with above-average investment advice play an influential role since they have large numbers of followers and retweets.

Our motivation to devise the diffusion hypothesis is twofold. First, considering a very short window for attention-induced retail investors to react to any news events, Twitter can be an important information transmission platform. In fact, most of the (breaking) news issued by news agencies is shared instantly via Twitter, thus making it an attractive platform for anyone to read news stories. Secondly, Twitter is one of the largest social networking sites. Therefore, CEOs, investment analysts, and bloggers can build their social media image by influencing users. Chen et al. (2019) suggest that information shared by CEOs and CFOs on Twitter helps to reduce information asymmetry in financial markets. Considering the virality and velocity of

Twitter, it should exacerbate the attention-induced buying behaviors of retail investors. Therefore, our diffusion hypothesis is as follows.

H2a: Information diffusion on Twitter positively affects the attention-induced buying behaviors of retail investors.

The unique characteristics of tweets, such as mentions, hashtags, retweets, and replies, make Twitter an obvious choice for a broad range of investors to discuss and share their opinions and is called salience of information signals (Huang et al., 2018; Li et al., 2019). Fiske and Taylor (2013) explain that salience has important implications for any stimulus to attract individuals. The salience of information signals not only increases the span of investors' attention on Twitter but also encourages them to participate in discussions on this platform. Cookson and Niessner (2019), using StockTwits, argue that interaction among investors on StockTwits results in disagreements, which are an important source of trading in the financial markets. In another study, Bollen et al. (2011) found that Twitter's mood can predict stock market returns. More recently, Baik et al. (2016) argued that local users share important information on Twitter, as compared to non-local users. Therefore, tweets' salience on Twitter plays an important role in highlighting information as compared to other social media platforms. We devise our final hypothesis to investigate the impact of tweets' salience on retail investors' trading behaviors.

H2b: Higher salience of information signals (Tweets) on Twitter is positively associated with the attention-induced buying behaviors of retail investors.

2.4. Sample Selection and Research design

2.4.1. Firm-level data

The sample period in our study is from Jan. 2012 to Dec. 2016. To eliminate the impact of survivorship bias and change to various index compositions, we selected all firms from the CRSP universe which met the following criteria: (1) the stock must be present in the TAQ master file; (2) the stock has a share code[12] of 10 or 11 in the CRSP database; 3) the stock has an average price at least greater than or equal to $3; (4) the stock has at least one year of trading data; (5) the stock has data available in CRSP, Compustat and I/B/E/S files; (6) the stock's ticker is matched with Twitter cashtags. We found a total of 2,675 firms that met these criteria.

[12] All the stocks with a share code of 10 and 11 are common stocks.

1.1.Market Microstructure Data

We construct net order flow (*NOF*) by classifying tick-by-tick millisecond transactions as buy/sell trades from daily trade and quote market microstructure data (*DTAQ*). As compared to any other proprietary data sets, the TAQ data set offers full market coverage at the millisecond level. It is offered at daily as well as monthly frequency levels. Within the TAQ data set, we use the Daily Trade and Quote (*DTAQ*) data set [13] as compared to Monthly Trade and Quote (*MTAQ*), as it has fewer errors while matching trades and quotes during the trades classification process (Holden & Jacobsen, 2014). Unlike previous studies, where researchers have used volume and stock returns as dependent variables, we use net order flow (NOF) as a dependent variable to understand investors' trading behaviors. The natural appeal of using *NOF* in this study stems from the fact that our hypotheses test the implications of *SMA* for trading behaviors, affecting either side of trades (buy/sell).

The rationale for using the market microstructure data is as follows. First, we need to classify trades into three groups to differentiate between retail and institutional investors. Therefore, transactional data from TAQ suffice for our needs, and following Lee and Ready (1991), we classify the trades as buy/sell. Second, the data set contains matched quotes from the National Best Bid and Offer (*NBBO*) file that is timestamped up to the millisecond level. The timestamp feature of these data makes it comparable to the instantaneous flow of tweets that are also timestamped in our data set. Third, the TAQ data set coverage is more than any other full-service brokerage firm data sets. For this study, we needed full coverage to avoid any self-selection bias in trading behaviors.

1.1.1. Trade Classification

Trades are classified as buy/sell trades to construct net order flow (NOF). In the first step, we delete all nonsensical cases when the market is crossed[14] or locked[15] since such cases can affect the trade classification process (Holden & Jacobsen, 2014). After cleaning the DTAQ data, we classify[16] every trade following the Lee and Ready (1991) method.[17] The LR algorithm

---

[13] We exclude quotes with non-normal quote conditions from DTAQ, e.g., A, B, H, O, R, and W. We also delete cases in which the bid of one exchange or market maker is greater than or equal to the same exchange or market marker. The quote condition must be normal, which excludes cases in which trading has been halted.
[14] A market is crossed when the best ask is strictly less than the best bid.
[15] A marked is locked when the best ask equals the best bid.
[16] Holden and Jacobsen (2014) extended market microstructure data by matching quotes up to the nanosecond level and then using the LR method. We replicate their approach and get the same results.
[17] Apart from the LR approach we also replicate the approaches of Ellis et al. (2000) and Chakrabarty et al. (2007). Our results and inferences remains the same.

classifies trades using the quote rule and tick rule. The quote rule classifies a trade as buyer initiated if the bid-ask midpoint is below the trade price and classifies a trade as seller initiated if the bid-ask midpoint is above the trade price, such as:

Same second quotes

$$trade\ k = \begin{cases} Buy_k, if\ P_k > M_t \\ Sell_k, if\ P_k < M_t \\ Use\ tick\ test\ if\ P_k = M_t \end{cases} \tag{1}$$

Previous second quotes

$$trade\ k = \begin{cases} Buy_k, if\ P_k > M_{t-1} \\ Sell_k, if\ P_k < M_{t-1} \\ Use\ tick\ test\ if\ P_k = M_{t-1} \end{cases} \tag{2}$$

The tick rule is adopted when the trade price and bid-ask midpoint are equal. Following the tick rule, a trade is buyer-initiated if the last executed trading price is lower than the current trading price. Similarly, a trade is seller initiated if the last executed trading price is higher than the current trading price. It is important to note that trades with equal bid-ask midpoints and current trading price equal to the last executed trade price cannot be classified as either seller or buyer initiated. Less than 1% of our sample trades are not classified since they do not meet the quote rule and tick test criteria.

2.4.2. Twitter Data

We have developed a customized programming application that connects with the Twitter application programming interface (API) and downloads tweets[18] that contain specific cashtags in our sample to collect data from Twitter. Since we have used Twitter cashtags as a direct proxy of investors' attention, we needed to exclude those tweets that carry multiple cashtags of different stocks. Therefore, every tweet in our sample is associated with only a single cashtag. To synchronize the Twitter data set with U.S capital markets, we converted the timestamps of tweets to Eastern Time (ET). Moreover, since a tweet is a collection of words,

[18] According to Twitter API documents, it is not possible to collect all tweets since every time a tweet search gives tweets from a different bucket. Given that, we replicated our programming application several times. During the final run, we saw a few additions in our tweet database as compared to the first run. We did not collect any deleted tweets, or any information blocked by the owners of tweets.

hashtags, cashtags, mentions, URLs, and media links, it was essential to parse[19] tweets to collect this information in our database. With reference to the ethical concerns involved in this study, we are fully aware that the Twitter API does not render any information that is protected by its users and requires their consent. Therefore, our Twitter data set only offers publicly available information.[20]

To further inspect the level of noise in our Twitter data set, we analyze the data set at various levels. (1) Although we only use common stocks in our sample, there are some instances where firm tickers are confused with their various share classes. In such cases,[21] Twitter users may use different cashtags for the same company. We test such cases using our cashtags validation approach. We drop cashtags that do not match our sample firms. (2) Some firms have tickers matching commonly used words and phrases. As our data collection was entirely based on Twitter cashtags, there was a possibility of encountering certain instances where users are writing these common words and phrases along with a dollar sign in their tweets and not intending to use a cashtag. For example, we came across several tweets that used the cashtag "$KIM." KIM is the ticker for Kimco Realty Corp.; however, we noted that fans of the celebrity Kim Kardashian also use this sign on various occasions after some inspection. Therefore, this validation check was important, and all such tweets were deleted. We also validated all the twitter cashtags with general meanings and a single alphabet, e.g., $GPS, $A, $B, and $ALL. Based on our cashtags validation approach, we identified tweets that were discussing firms' related information. Da et al. (2011) tag some tickers as noisy tickers and conduct additional analysis to check their results' validity. We extend their approach and add further tests to check the validity of our Twitter sample.

2.5. Variables Construction

2.5.1. Net Order Flow

A review of the relevant literature suggests that there seems to be a consensus that the segregation of trades into groups usually clarifies and highlights the differences between retail and institutional investors. However, it is important to note that small trades turnover has increased rapidly since 2000, whereas the turnover of large trades has remained constant. One

[19] For tweet parsing, we use python 3.6 and Natural Language Toolkit (NLTK) 3.4. Parsers can be cloned from our GitHub page https://github.com/Mustabsar

[20] To ensure that tweets are downloaded anonymously, we use one-way encryption (hashing algorithm) which assigns a unique value to every tweet and every user and cannot be unencrypted. This ensures that any tweet/ user-related information cannot be saved in our database.

[21] CRSP permnos matching also helps us identify such cases in trading data.

of the major reasons for this trend in small trades is that institutional investors tend to break down their large orders into small orders to reduce transaction costs (Barber et al., 2008; Hvidkjaer, 2008). In contrast, Yuan (2015) argues that it is unlikely that such a structural break in the institutional order flow coincides with a change in how individual investors allocate their attention.

For the purposes of this study, trades are segregated into three groups by the market value of trades. Small trades are trades up to \$10,000; medium trades are more than \$10,000 and up to \$50,000; and large trades are more than \$50,000. Trade groups were adjusted for inflation and are based on the 1991 dollar rate and using a consumer price index. Overall, our trades segregation method is in line with Lee and Radhakrishna (2000), who used a \$10,000 trade size as the threshold for retail investors and trades greater than or equal to \$50,000 as the threshold for institutional investors. In another study, Barber et al. (2008) divided trades into quintiles, with the lowest being retail trades and the highest being institutional trades. Following Yuan (2008), we calculate net order flow (*NOF*) for each trade group as follows:

$$NOF = \frac{\sum_{i=n}^{t} MV_{Buy} - \sum_{i=n}^{t} MV_{Sell}}{\sum_{i=n}^{t-1} MV_{MKT}} \tag{3}$$

Where MV_{Buy} is the market value of buy trades, MV_{Sell} is the market value of sell trades and MV_{MKT} is the lagged market value of NYSE and AMEX.

2.5.2. Abnormal Returns

To understand the impact of price pressure, we used 5x5 book-to-market portfolio returns to create an abnormal return variable.

$$Abnormal\ Return = R_{i,t} - R_{m,t} \tag{4}$$

Where $R_{i,t}$ is the return of stock *i* at time *t* and $R_{m,t}$ is the benchmark return of the market *m* at time *t*.

2.5.3. Social Media Attention

To calculate social media attention (SMA), we aggregate the total number of tweets for each firm on a daily basis and measure firm-day Twitter activity. As discussed earlier, to match our Twitter sample with the TAQ data set, we divided tweets into groups based on the trading and non-trading hours of U.S financial markets. The trading hour group is matched with the TAQ data set, and for the non-trading data set, we use time lag for pre- and post-market periods. Whereas the pre-market period is the time group from 12 AM to 09:29:59 AM eastern time (before the markets open), trading hours are regular trading hours when the markets are open; the post-market period is the time group from 4 PM – 11:59:59 PM eastern time (when the markets are closed). Our *SMA* variable is as follows:

$$SMA = \ln\left(1 + Tweets_{i,t}\right) \tag{5}$$

Where *t* is the time at which a tweet is posted; to ensure that SMA is not affected by any time trends, SMA is detrended using the day of the week, week of the month, and month of the year.

2.5.4. Salience of Tweets

Salience is defined as the attention-grabbing characteristics of information signals (Huang et al., 2018). For this purpose, we developed a text parser using a natural language toolkit (*NLTK*) to extract Twitter Hashtags,[22] Twitter Mentions,[23] Tweets with media,[24] and Tweets with URLs.[25] While using the Twitter API service, we also harvested user verified status, geographic locations[26] (if given), user followers and following, and date of joining Twitter. It is pertinent to note that Twitter also allows its users to verify their Twitter accounts, but this is not a mandatory feature. In our Twitter data set, only 3% of users have verified status. Although most users can change their geographic location at any time, some of them prefer not to mention it. However, Takhteyev et al. (2012) recommend that irrespective of its reliability, users' location can offer some valuable insights into Twitter users' geography. To understand the impact of information diffusion, we calculated tweet reach as the sum of distinct users' followers who posted firm-specific tweets. Tweets posted by market influencers and

[22] Hashtags is a metadata function that enable users to sift through current trends as well contribute to discussions by using # with trending words.
[23] A mention is a reference to another user on Twitter. Usually users use mentions to reply or refer to another user in a discussion thread on Twitter.
[24] Twitter users can share images and videos within tweets.
[25] These links are shared by users to direct readers to third party applications or websites for further details. Our text parser cannot encode encrypted URLs within tweets. Hence, we are only able to parse unencrypted URLs.
[26] We developed a program using Google Geotagging API and matched every user location with it. The program saves verified locations along with longitude and latitude. Not all users write their location and not all users make it public. Therefore, we only collect the public information of users.

carrying any material information have higher tweet reach, more retweets, and lengthier discussion threads (replies). Overall, we capture almost every aspect of the salience of firm-specific tweets.

2.6. Baseline Model and Specifications

Our data set is characterized as longitudinal data since it has both time series and cross-sectional dimensions. Therefore, we create a panel data set for the reason that it can deal with firm-specific heterogeneity as well as unobservable factors. Consequently, we get an unbalanced panel data set because not all firms have small, medium, and large trades on each trading day. However, it is unlikely that an unbalanced panel will have any effect on our regression results. Our baseline model is as follows:

$$NOF_{i,t} = \beta_0 + \beta_1 SMA_{i,t} + \phi Controls_{i,t} + V_i + V_t + V_p + \varepsilon_{i,t} \tag{6}$$

Where i represent firms in our sample and t represents trading days. $NOF_{i,t}$ is the net order flow of the stocks of firm *i* at time *t,* for small, medium, and large trades, $SMA_{i,t}$ is the social media attention of firm *i* at time *t*. $Controls_{i,t}$ is a set of control variables. We estimate the equation using a fixed-effects estimator to account for unobserved firm-specific heterogeneity (V_i). We also control for time (V_t) and industry (V_p) fixed effects by including month and industry dummies, capturing time-varying and industry-specific movements. Following Petersen (2009), we cluster standard errors at the firm level to control for within-group correlation.

We use several control variables in our regression to validate our research model. Following Chordia and Subrahmanyam (2004), we used one-day lagged *NOF* and the previous year's abnormal returns to control long- and short-term economic specifications. To account for the traditional proxies of attention, we use abnormal turnover and absolute abnormal returns to control for extreme trading activity and returns (Barber & Odean, 2008; Da et al., 2011). To control for institutional attention, we use Bloomberg's institutional attention data, and following Ben-Rephael et al. (2017), we measure abnormal institutional attention (*AIA*). Da et al. (2011) used Google search volume as a proxy for retail investors' attention. Therefore, to control for retail investors' attention, we measure the abnormal search volume index (*ASVI*).

Hirshleifer and Teoh (2003) find that higher analyst coverage improves market efficiency and supports investors by providing firm- and market-level information. Therefore, we control for analyst coverage. We also control for stock-specific characteristics such as price volatility, and for firm-specific attributes such as firm age, firm size, book-to-market ratio, and retail ownership. These control variables are consistent with previous studies by Cookson and Niessner (2019) and Giannini et al. (2019), among others.

We also control for firm-level media coverage using the Dow Jones news service and significant news developments. Fang and Peress (2009) used Dow Jones news coverage to investigate the impact of media coverage on the cross-section of returns. Similarly, significant news developments are offered by Thomson Reuter's news service. These stories are usually flash news informing the market about any significant developments that could affect financial markets. Moreover, this is just-in-time news that is commonly shared by the Twitter community.

2.7. Empirical Analysis

2.7.1. Summary Statistics

Our decision to use Twitter activity during trading hours, as a standard period to aggregate tweets, is based on Twitter activity during a 24-hour period. The most Twitter activity takes place during trading hours. Similarly, days of the week Twitter activity suggests that most tweets are posted on weekdays, compared to weekends. The average number of tweets posted during trading hours and a 24-hour period is 5.77 and 11.99, respectively. The average number of active days per stock based on Twitter activity is 754 and 932 for trading hours and 24-hour periods, respectively. Taking advantage of the breadth of information we have in Twitter's data, we also summarise Twitter activity at the user level. Therefore, the average number of tweets per user in our Twitter sample is 8.3 and 6.6 for trading hours and a 24-hour period, respectively.

[Insert Table – 2.1 here]

[Insert Figure – 2.1 here]

Twitter users' experience in our sample suggests that half of the users have at least 2.5 years of Twitter experience, with an average of 2.65 and 2.48 years during trading hours and a 24-hour period, respectively. Our user sample during trading hours presents that only 2.98% of users have verified their Twitter accounts, as compared to 2.34% of users during a 24-hour period. This further suggests that most verified users prefer to post during trading hours. Finally, we analyze users' geographic locations. Although quite subjective, this analysis offers some useful insights into the geographic distribution of tweets. We divide users into three groups: users with unspecified locations, users from the USA, and users from the rest of the world (*ROW*). The geographic distribution of tweets suggests that 30.78% of users are from the USA, 12.76% of users are from ROW, and 56.46% of users have unspecified locations during trading hours. In the case of a 24-hour period, 27.79% of users are from the USA, 12.85% of users are from ROW, and 59.36% of users have unspecified locations.

Table 2.2 presents summary statistics for financial data. Panel A shows firm-level data for each trade group. *NOF* is 0.16 in the small trades group compared to 0.05 in the medium and 0.06 in the large trades groups. The higher mean value of NOF suggests that investors prefer to place small trade orders. Furthermore, it is noted that small trade orders are more frequent for smaller firms, i.e., an average firm size of 7.59 bn US dollars as compared to medium and large trade groups. On average, firms in the small trades group are young (22.88 years), have less analyst coverage (6.36), a higher book-to-market ratio (0.47), and higher retail ownership (0.31) as compared to the medium and large trades groups. The differences between the summary statistics of all the trade groups are statistically significant.

Since we group trades into small, medium, and large to differentiate between retail and institutional investors, we further investigate these groups' validity by comparing them with the existing literature. Overall, there is a consensus in the literature that trade size is an important proxy for different types of investors in financial markets (Barber et al., 2008; Shanthikumar, 2012). Although, Hvidkjaer (2006) argues that the exponential increase in small trades since 2001 is due to decimalization and the breaking down of large trades into smaller ones, it is unlikely that this change has any significant impact on attention-induced buying behaviors.

[Insert Table – 2.2 here]

Our trade classification data set is consistent with Barber et al. (2008) and Yuan (2015). Da et al. (2011) used Dash-5 monthly reports to mimic individual investors' trading behaviors. Unfortunately, that data set is not available for our sample period. However, the average volume in our small, medium and large trade groups falls within the same threshold of Dash-5 monthly reports. Panel B presents trade data by trade-group and classification. Overall, the average volume of small trades is 116.94 shares with an average trade value of $3,537.98; the average volume of medium trades is 2,152.37 shares with an average trade value of $47,249, and the average volume of large trades is 14,492.14 shares with an average trade value of $498,950. These statistics further present the validity of our trade groups and trade classifications in line with recent literature.

2.7.2. Net Order Flow and Social Media Attention

An important question is whose attention does SMA capture? Since Twitter can diffuse information quickly, the real-time impact of information shared on Twitter should be greater during trading hours. Hence, the natural instinct is to test our hypotheses for SMA during trading hours. Table 2.3 presents contemporaneous fixed effect regression results on small, medium, and large trade groups. Model (1) indicates that SMA has a positive significant impact on the NOF of small trades, which further suggests that retail investors are net buyers of attention-grabbing stocks. We control previous day order flow and returns for the previous year following Chordia and Subrahmanyam (2004); this helps us control for economic information in both the short- and long run. The instantaneous flow of information on Twitter can be affected by extreme returns and abnormal turnover during trading hours. Therefore, we also control for return and volume shocks in our model. To test further for the presence of retail investors, we estimate our baseline model in the presence of Firm Age, Analyst Coverage, and Firm Size as additional controls in Model (2). We are also interested in understanding the role of retail investors, as in Fang and Peress (2009), and price volatility since Twitter is an important platform for discussion where users can agree or disagree with each other's opinions and disagreements and a divergence of opinions may increase price volatility (Cookson & Niessner, 2019; Giannini et al., 2019). Model (3) presents results after adding Book-to-Market ratio, Retail Ownership, and Price Volatility to our model.

[Insert Table – 2.3 here]

We find significant negative results for firm size in small trades groups as compared to an insignificant relationship in large trade groups during trading hours. Our results remain consistent and provide further evidence that retail investors trade in small firms with less analyst coverage and are comparatively younger. It is pertinent to note that firms with less analyst coverage may have less information available in the market, consequently diverting investors' attention towards alternative channels of information, i.e., Twitter. Furthermore, there is more retail ownership of firms in the small trades group, which is positively significant compared to no relationship in the large trades group. Our results are consistent with Darrough and Stoughton (1990), who use Dash-5 reports as a proxy for trades by retail investors, and Antweiler and Frank (2004), who use the number of trades from the DTAQ data set as a proxy for small trades.

While taking into account the real-time impact of SMA, it is vital to investigate the impact of SMA on the next day's NOF. Table 2.4 presents the results from a time sequencing regression between SMA and next day NOF. Interestingly, the results suggest that the positive association between SMA and NOF is reversed on the following days with an increasingly negative impact during the rest of the week. This result suggests that investors may overestimate their understanding of the information environment because of SMA and are willing to take higher risks, thus overlooking details (Hirshleifer & Teoh, 2003). Those who make investment decisions later realise the consequences of their shoddy analysis and start selling. Barber and Odean (2000) present compelling evidence that, overall, retail investors earn less than market returns, and the increase in trading activity is due to retail investors' overconfidence.

[Insert Table – 2.4 here]

2.7.3. Social Media Attention and Media Coverage

Twitter is a popular conduit for information transmission because of its salience features. Users can share anything, including images and videos, which can add to the distinctive features of their tweets to attract larger audiences (Lee et al., 2015; Rainie et al., 2012). A recent strand of research supports the argument that Twitter has transformed the news industry into ambient journalism, where every journalist mainly relies on digital systems such as Twitter to collect and share news via their Twitter handles (Hermida, 2010). Motivated by the role of videos and images shared by Twitter users in tweets with cashtags, we developed another variable, has_media, a dummy variable that equals one if a tweet contains any video or image, otherwise 0. The general understanding is that tweets with any kind of media can get more attention and likes as compared to tweets with no media content (Nekrasov et al., 2021). In addition, we also test for the implications of flash news released during trading hours since investors pay attention to flash news as they happen. We use the Sig. Developments variable that covers all the important news related to financial markets. Interestingly, most flash news is released by news agencies via their Twitter handles.[27] Finally, to account for firms' overall

[27] Bloomberg started its Twitter handle @tictoc to share flash news as it happens in February 2011.

media coverage, we develop another variable, News Coverage DJ, by aggregating the Dow Jones news coverage of the firm daily.

Table 2.5 presents the regression results for net order flow, social media attention, and news coverage. For every trade group, we run three regressions. In all three regressions, we use has_media to control for tweets' visual impact. Interestingly, our results suggest that Twitter's visual impact remains consistent even in the presence of other media variables. In Model (1), Sig. Developments and News Coverage DJ are positively significant. Interestingly, the coefficient of Sig. Developments is four times larger than the coefficient of News Coverage DJ. To test the implications of Sig. Developments on SMA we interact SMA with Sig. Developments in Model (2). The results suggest that flash news plays an important role in attracting investors via Twitter. These findings are consistent with Kwak et al. (2010), who found that Twitter users tend to share news headlines which further diffuse at a higher rate. In Model (3), we interact SMA with News Coverage DJ. The coefficient of interaction between SMA and News Coverage DJ remains smaller compared to Dow Jones news. These findings suggest that just-in-time news attracts more investors' attention on Twitter as compared to detailed news stories.

[Insert Table –2.5 here]

Across the group, a comparison of all the regression results between small, medium, and large trades provides compelling evidence that retail investors are those whose trading behaviors are affected by SMA and other media variables. Interestingly, our *has_media* variable suggests that tweets' visual impact remains consistent across all three trade groups. However, the magnitude of its coefficient is smaller in large trades as compared to small trades. However, the rest of the media variables have no significant impact on large trades, suggesting that institutional investors consume information through various other sources (Ben-Rephael et al., 2017); and even then, most of them can predict news (Hendershott et al., 2015). It is pertinent to mention that the signs for the rest of the control variables remain consistent across all the trade groups and in all the regressions. Therefore, our findings are consistent in finding that media coverage exacerbates the impact of *SMA* on *NOF* for retail investors and, as a result, retail investors experience attention-induced buying behavior.

2.7.4. Social Media Attention and Price Pressures

To examine the impact of attention-induced buying behaviors on stock prices, we use 5x5 book-to-market and firm size portfolio returns as our dependent variables. Since firm returns are cross-sectionally correlated, there is a probability of underestimating standard errors or overstating levels of significance. Therefore, to deal with this problem, we use two-stage Fama and MacBeth (1973) regression. However, in a panel data set, cross-sectional and time-series dependence can also affect our results. Therefore, to handle these forms of dependence, we use Newey and West (1987) corrected Fama-McBeth standard errors with four lags (Petersen, 2009). Our basic regression model is as follows:

$$AR_{i,t} = \beta_0 + \beta_1 SMA_{i,t} + \phi Controls_{i,t} + \varepsilon_{i,t} \tag{7}$$

Where $AR_{i,t}$ is the abnormal return of firm i at time t. It is pertinent to note that in this model we use *ASVI, Firm Size*, BSI_{Retail} and *Retail Ownership* to proxy for retail investors. Where BSI_{Retail} is the Buy-Sell imbalance of retail investors.

Table 2.6 presents the Fama-McBeth regression results for SMA and price pressures. We use demeaned cross-sectional variables that are standardised to account for unit variance except for *Abnormal Returns* which are reported as basis points. Instead of using trade groups, we add proxy variables for retail investors following Da et al. (2011). In Model (1), we estimate the regression using Equation (5). We find a positive association between *SMA* and *Abnormal Returns*. Interestingly, the impact of *SMA* is stronger than *ASVI,* suggesting that *SMA* is a real-time proxy for retail investors' attention. The results also support *SMA* and *ASVI* being completely different proxies for attention, as they measure attention in an entirely different fashion from each other.

In Model (2), we add BSI_{Retail} and *Retail Ownership* as proxies for retail investors. BSI_{Retail} is the daily difference between the buy and sell volumes of retail trades. Therefore, we use BSI_{Retail} as a proxy for retail investors' trading volume. If there is more trading activity by retail investors due to attention-induced buying behavior, then the coefficient of BSI should be positive, and we expect the same for *Retail Ownership*. Interestingly, our regression results tell the same story. Both BSI_{Retail} and *Retail Ownership* has a significant positive association with *Abnormal Returns* in the presence of *SMA*, suggesting that short-term price pressures are due to the attention-induced buying behaviors of retail investors.

[Insert Table – 2.6 here]

In Models (3) to (5), we interact SMA with *Firm Size*, BSI_{Retail} and *Retail Ownership*, respectively. The coefficient of interaction between *SMA* and *Firm Size* is negative, thus offering evidence that increases in price pressure are due to *SMA,* especially in small firms. Moreover, the coefficients of the interaction of BSI_{Retail} and *Retail Ownership* is positive, further indicating that higher trading volumes are due to retail investors' attention-induced buying behaviors. Overall, our results are consistent with the existing strand of research by Da et al. (2011), Giannini et al. (2019), and Sprenger et al. (2014). Our findings in this section support the evidence from the previous section that retail investors are net buyers of attention-grabbing stocks and this attention-induced buying behavior results in price pressures.

2.7.5. The salience of firm-level tweets and Net Order Flow

This section examines the moderating role of the salience of firm-level tweets on net order flow (*NOF*). As discussed in the previous section, several distinctive features of tweets transmit meaningful information signals to market participants. We divide the salience of firm-level tweets into two groups. First, user-level salience makes Twitter a large social network that can diffuse information faster than any other information channel, i.e., *Tweets Reach* and *Users' Concentration*. Second, tweet-level salience plays a pivotal role in attracting users' attention based on several distinctive features of a tweet—for example, the verification status of users, replies, and retweets, i.e., *Tweet Characteristics*.

2.7.6. Tweet Reach and Users' Concentration

We can distinguish between less influential and more influential Twitter users by counting the number of followers any users have on their Twitter profiles. Influential users have millions of followers. Consequently, a single tweet posted by an influential user can reach millions of followers and increase the impact of that tweet. To win the race of top influencers, Twitter users keep posting enticing content that can attract many users, who consequently ending up following them (Cha et al., 2010). Twitter is one of the largest social media platforms, and it offers unrestricted access to content shared by its users. For this reason, several CEOs and investment gurus keep sharing their views[28] about specific topics, which are sometimes related to their firms. We define *Tweet Reach* as the sum of distinct users' followers who posted tweets with cashtags. The general understanding is that information that is

[28] Carl C. Icahan is an American businessman and investor. He has more than 366,000 followers. He shares popular stock commentaries and updates with investors related to his companies and investments in financial markets. Guy Adami has more than 281,000 followers and a verified Twitter account. In his tweets, he offers commentary on stock markets and useful insights into the financial performance of listed and non-listed companies.

disseminated to a large social network may have a larger influence on retail investors' attention-induced buying behaviors.

In Table 2.7, we divide *Tweet Reach* into terciles, ranging from low to high Tweet Reach in each trade group. We estimate our regression model using *NOF* as a dependent variable on all the terciles in each trade group. We find that the influence of *Tweet Reach* increases when we move from a small social network to a large one. Most importantly, the coefficients of these results remain positive and significant for small trades with an incremental effect of Tweet Reach on the magnitude of the coefficients from low to high terciles. However, the results change for large trades where we find no significant evidence that *Tweet Reach* has any influence on institutional investors. Our results support the hypothesis that a large social network has the ability to influence the attention-induced buying behaviors of retail investors. Moreover, the results remain robust in the presence of Sig. Development and News Coverage DJ, as well as firm-level and stock-level controls.

[Insert Table – 2.7 here]

We now discuss the role of users' concentration and whether it influences the buying behaviors of retail investors. Users' concentration is defined as the total number of unique users posting firm-specific tweets. The greater the number of users posting tweets with certain cashtags, the larger the social network and the higher the heterogeneity of users on Twitter. For example, a cashtag being posted by a large number of users means the Twitter crowd is already paying attention. Conversely, a tweet with a cashtag that can reach millions of users means the Twitter crowd can actually have access to the same information. Table 2.8 divides users' concentration into terciles, ranging from low to high users' concentration. We find a positive association between NOF and SMA for all the terciles in the small trades group and no association in the large trades group. More importantly, like Tweet Reach, we find an incremental effect of users' concentration, suggesting that an increase in the number of users posting firm-specific tweets increases the attention-induced buying behaviors of retail investors.

[Insert Table – 2.8 here]

2.7.7. Tweet Characteristics

Twitter users can share any tweet and can create threads on different topics for discussion purposes. Similarly, Twitter allows celebrities and influencers to verify their profiles to increase the credibility of their tweets. In this section, we present our regression results based on tweet characteristics. For this purpose, we segregate users based on verified user profiles. We estimate our regressions on the trade groups of verified and non-verified users. The general instinct is that tweets posted by verified users may have a larger influence as compared to tweets posted by non-verified users since the credibility of users is an important issue in social and conventional media (Dyck et al., 2008). Our regression results in Table 2.9 and Panel A are consistent with existing evidence and suggest that verified users' tweets have a larger influence on retail investors' trading behaviors than those of non-verified users. It is pertinent to mention that such features are not available in Google search; therefore, the ASVI variable is significant, whether it be for small, medium, or large trade groups.

[Insert Table – 2.9 here]

There are different ways of paying attention to tweets. For example, retweeting or replying to a tweet increases its importance as compared to those tweets which have no retweets or replies. In our Twitter data set, we also harvest the number of retweets and replies a single tweet gets after it is posted. We divide our trade groups based on replies and retweets. Our results in Table 2.9, Panels B and C, provide compelling evidence in favor of replies and retweets, respectively, suggesting that tweets that have more discussions and retweets have a larger influence on retail investors' buying behaviors. These findings are consistent with Chen et al. (2014), who used data from SeekingAlpha and conclude that comments on any social media posts reiterate the importance of those posts. It is pertinent to mention that our SMA coefficient is significant for large trades by non-verified users, with no replies or retweets, which further suggests that Twitter users discuss a wide variety of topics, including trades by institutional investors.

2.8. Robustness and External Validity

To the best of our knowledge, this is the first study to use Twitter data to study their influence on retail investors' trading behaviors. Therefore, we cannot compare our results with other studies. However, we conduct several robustness checks to validate the findings of this study.

2.8.1. SMA and alternative proxies for Attention

Since we have developed our argument based on the real-time influence of SMA on retail investors' trading behaviors, we must compare existing proxies for attention with SMA.

For example, Da et al. (2011) used ASVI as a direct proxy for the attention of retail investors and concluded that it predicts market returns. Ben-Rephael et al. (2017) use AIA as a proxy for institutional attention to news in Bloomberg terminals and confirm that AIA measures institutional attention, which is different from ASVI and leads to retail attention. Other traditional proxies for attention such as firm size, analyst coverage, media coverage, extreme returns, and abnormal turnover have been used extensively by researchers to investigate their impact on retail investors' trading behaviors. Based on the current line of research, we regress SMA using alternative proxies for attention as our explanatory variables.

[Insert Table – 2.10 here]

The results in Table 2.10 present that *SMA* is an entirely different measure of attention as compared to alternative proxies for attention. Our results imply that *ASVI* does not explain *SMA,* suggesting that they are both different proxies for attention that capture investors' attention in an entirely different fashion. Interestingly, a lot of variation in *SMA* is explained by media variables, which further supports our finding that *SMA* is used for the transmission of information instead of being a primary source of information. *AIA* has a positive coefficient suggesting that *SMA* and *AIA* are both used for information consumption by different sets of market participants, i.e., institutional investors and retail investors. *Analyst coverage* has a negative coefficient which further suggests that firms covered less by analysts are mostly discussed in alternative sources of information. Therefore, based on the evidence, we can confirm that SMA is a unique proxy for attention and is different from *ASVI* and *AIA,* as well as traditional proxies for attention.

2.8.2. SMA and non-trading hours

Although we have already discussed that most tweets are posted during trading hours, it is imperative to understand the influence of SMA during pre-and post-market periods. Our motivation is twofold to test these market settings. First, if SMA has a real-time impact during trading hours, tweets posted during non-trading hours should either be insignificant or have less predictive power than trading hours. Second, Engelberg et al. (2012) suggest that some famous investment gurus share their commentaries once markets are closed, and they can predict the next day's market returns. Table 2.11 presents the regression results for SMA during non-trading hours, suggesting that SMA during non-trading hours can predict the attention-induced buying behavior of retail investors. However, the magnitude of the SMA coefficient is smaller than trading hours. There is less influence of SMA during non-trading hours for various reasons. For example, during non-trading hours, the wisdom of the crowd becomes less accurate due to smaller social networks and lower diversity than during trading hours (*see Figure 1*). Moreover, the post-market period is a crucial period for traders, especially for retail investors. During this time, traders try to watch stock market commentaries, analyse their investments, or compare their *"Do it Yourself"* analyses (Chen et al., 2014). Thus, every piece of information related to stocks is vital. Overall, the evidence from non-trading hours validates our previous findings that SMA's real-time implications are greater than non-trading hours.

[Insert Table – 2.11 here]

2.8.3. Ranking Analysis of Tweets

It is imperative to understand the content of tweets that have been used to measure SMA in this study. As discussed earlier, we use a natural language toolkit (NLTK) to analyse the text of tweets. For the purposes of this analysis, we parse tweets to save cashtags, mentions, and hashtags in our database and rank them based on their frequency in our data set. Table 2.12 Panel A presents a list of the Top 50 discussed firms. To the best of our knowledge, this is the first study to offer US Firms Rankings based on Twitter cashtags; therefore, we cannot compare our ranking list with any other study. However, Cookson and Niessner (2019), using a sample from stocktwits.com, ranked the top-100 firms based on cashtags. The Top 100 ranking list by Cookson and Niessner (2019) is matched with 70% of firms in our Top 100 ranking list, and the rest of them are matched with Top 250 firms. Since these two social media platforms are not entirely similar, a difference in trends is quite natural.

[Insert Table – 2.12 here]

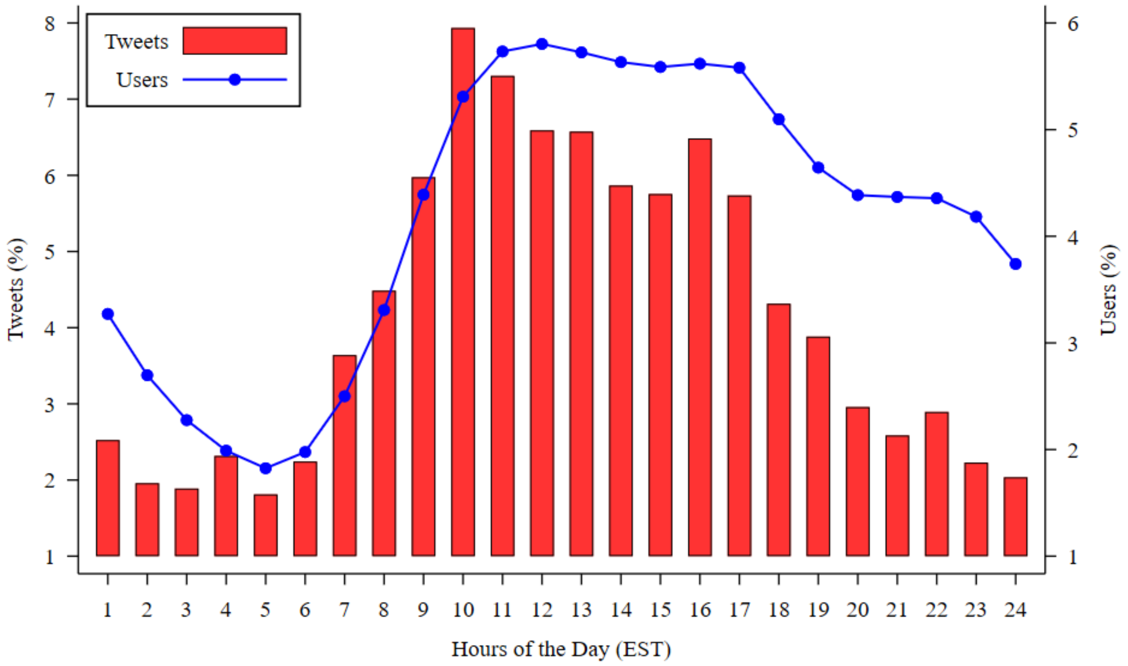
Panel B presents a list of the Top 15 mentions in our Twitter sample in order of frequency. All the top mentions in our sample either post a financial tweet or share/ discuss any stock on Twitter using cashtags. In our top mentions list, @jimcramer is the number one mention. Panel C presents a list of the Top 15 hashtags in our Twitter sample. It is pertinent to mention that more than 95% of hashtags concern finance-related topics.

2.9. Chapter Conclusion

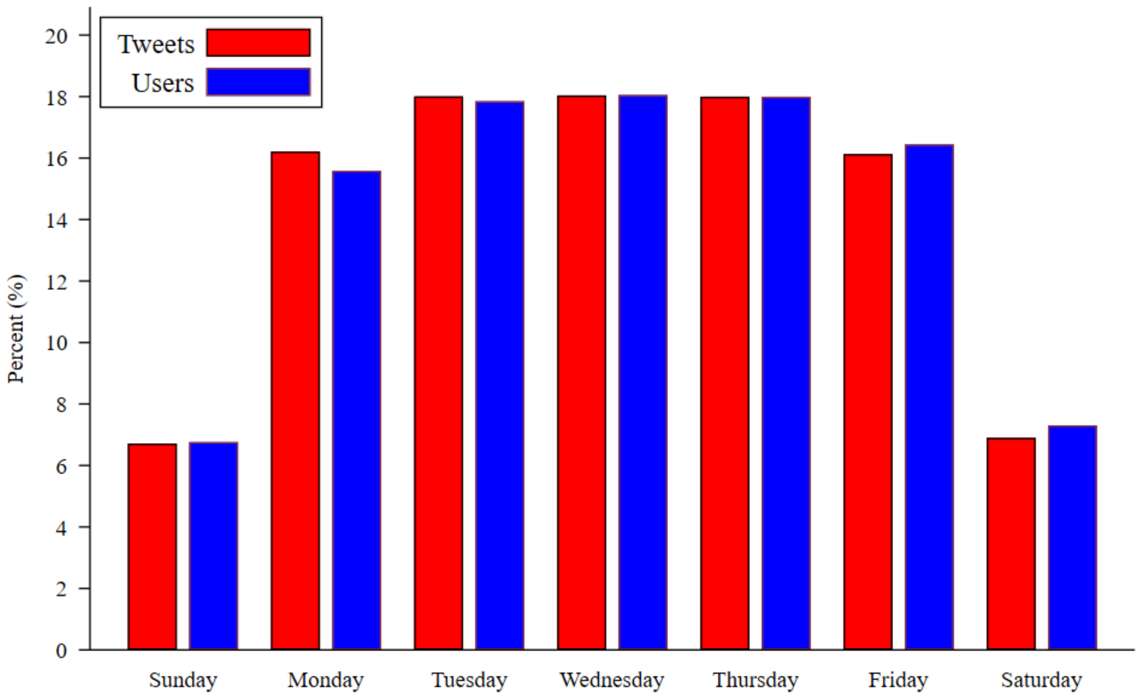
Social media have transformed the way investors share information in financial markets. Using a novel data set of more than 32 million tweets and a five years of market microstructure data set, this study offers some useful insights. Our battery of experiments suggests that tweets with cashtags can be used as a direct proxy for retail investors' attention, which is entirely different from alternative proxies for attention. Moreover, further tests on Twitter heterogeneity suggest that SMA becomes more important in an extensive social network of millions of users where it has more implications for retail investors. SMA remains insignificant for institutional investors. Segregating social media data set into pre-market, trading hours, and post-market periods also offers concrete evidence that SMA has a real-time impact on individual investors' trading behaviors as compared to pre-market and post-market periods. Analysing the post-market period, SMA reveals that investors are indifferent during this period, and overall the significance of SMA is significantly reduced as compared to trading hours. Similarly, the pre-market period is a quiet period for investors since there is the least number of user engagements on Twitter during this period. Overall, our results support the *attention hypothesis* that retail investors are net buyers of attention-grabbing stocks. In addition, our results contribute to the emerging scholarly work on social media attention and its implications for financial markets. This paper has several implications for various stakeholders in financial markets.

Our findings suggest that retail investors are one of the most vulnerable groups in financial markets in terms of information acquisition from various available resources generally and Twitter specifically. Unlike any blog site, social media platforms like Twitter need to be regulated to reduce noise among the already available abundance of information. Although the Securities and Exchange Commission has encouraged companies to disseminate information using social media platforms, this needs to be adequately enforced. For example, in our sample of 2,675 companies, only 25% of companies have Twitter accounts, and of these, only 9% of companies share their financial information with investors using these platforms. The absence of a reliable source for information dissemination can be harmful to investors and markets overall. While analysing the salience of tweets, a list of the top 15 Twitter mentions offers some critical insights. First, there is an emerging tendency of users to follow investment gurus who have millions of followers, irrespective of understanding the implications of their investment strategies. Second, it shows that retail investors rely more on DIY analysis, which can help them to allocate their funds as efficiently as possible. In both cases, social media facilitate a reduction in information asymmetry.

Figure 2.1: Distribution of Tweets and Users at different time intervals



Panel A: Hourly distribution



Panel B: Daily distribution

Notes: The figure presents the distribution of Tweets and Users at different time intervals. Panel A shows the distribution within the 24 hours period. Timestamps are converted from the universal time zone to eastern time. Panel B shows the distribution of Tweets and Users on days of the week.

Table 2.1: Summary Statistics of Twitter data

Notes: The table presents the summary statistics of firm-level and user-level Twitter data. Panel A presents firm-level Twitter data. ***Tweets*** are the number of firm-specific tweets; ***Users*** is the number of users who post firm-specific tweets; ***Tweets Reach*** is the sum of the distinct users' followers who post firm-specific tweets; ***Hashtags*** and ***Mentions*** are the numbers of hashtags and mentions parsed from the firm-specific tweets; ***Active Days per Stock*** report the number of trading days per stock when there is at least one firm-specific tweet. Panel B presents user-level data of the firm-specific tweets. ***Followers*** and ***Following*** represent the number of users who follow the user who posted the firm-specific tweet and the number of users being followed by the user who posted the firm-specific tweet, respectively. ***User Experience*** is the number of years since the user joined Twitter and posted the first firm-specific tweet in our sample; ***Tweets per User*** count the daily number of firm-specific tweets posted by a single user; ***Active Days per User*** counts the total number of days when a user has posted at least one firm-specific tweet. ***Likes*** are the total number of likes a users' tweets received since posted on Twitter. ***Lists*** are generated to view the timelines of the specific accounts. ***Verified*** presents the number of verified Twitter accounts, and ***Non-Verified*** presents the vice versa. Users' locations are identified from public profiles of the users. The number of trading days is 1258, and the total number of days (all-inclusive) is 1825. The timestamp on tweets is converted from Universal Time Coordinated (UTC) to Eastern Time (E.T.). Trading hours are when Wall Street is active (09:30 AM – 4 PM ET).

Panel A: Firm-level Twitter data

Description	**Trading Hours**						**24 Hours**					
	Mean	**SD**	**P25**	**P50**	**P75**	**Max**	**Mean**	**SD**	**P25**	**P50**	**P75**	**Max**
Tweets	5.77	31.03	0	1	4	167,786	11.99	57.58	1	3	8	17,667
Users	4.05	16.64	0	1	3	8,130	7.48	25.69	1	3	7	10,203
Tweets Reach	61.34	588	0	0.79	13.66	189,237	99.99	838.63	0.03	5.97	34.65	252,919
Hashtags	2.82	19.96	0	0	1	5,625	7.29	51.49	0	0	3	9,996
Mentions	0.87	10.07	0	0	0	7,532	1.54	16	0	0	0	10,696
Active Days per Stock	754	294	516	721	999	1,258	932	242	761	953	1140	1,258

Panel B: Twitter's User-level data

Description	**Trading Hours**						**24 Hours**					
	Mean	**SD**	**P25**	**P50**	**P75**	**Max**	**Mean**	**SD**	**P25**	**P50**	**P75**	**Max**
Users' Followers	4,876.66	22,4042.70	111	401	1173	71,413,125	3,982.97	186,417.40	125	409	1,074	71,413,125
Users' Following	1,122.03	8,745.44	138	375	884	3,062,635	1,002.55	6,989.24	150	371	828	3,062,635
User Experience	2.65	1.92	1	2.42	3.92	10	2.48	1.85	0.92	2.17	3.67	10
Tweets per User	8.3	125.63	1	1	2	33220	6.6	143.04	1	4	2	50,089
Active Days per User	2.35	15.01	1	1	1	1656	2.26	14.09	1	1	1	1,523
Number of Likes	2,208.97	11,792.26	0	1	366	1,896,122	1,905.69	10,700.56	0	0	128	1,896,122
Number of Lists	65.65	1,904.12	1	5	27	1,196,651	52.76	3,073.94	1	4	18	2,386,905
User Status	**Verified**			**Non-Verified**			**Verified**			**Non-Verified**		
	2.98%			97.20%			2.34%			97.66%		
User with Locations	**Unspecified**		**USA**		**Rest of the World**		**Unspecified**		**USA**		**Rest of the World**	
	56.46%		30.78%		12.76%		59.36%		27.79%		12.85%	

Table 2.2: Summary Statistics of Financial data

Notes: The table presents the summary statistics of firm-level financial data and aggregate trading data for each trade group. Panel A presents firm-level financial data for each trade group. ***Net Order Flow*** is measured as the difference between buyer and seller-initiated trades normalized by the NASDAQ-NYSE daily market capitalization. ***Abnormal Turnover*** is calculated by dividing the current volume by the previous year's volume. ***Firm Size*** is the quarterly market value of equity. ***Firms' Age*** is the number of years since their data is first reported in the Compustat database. ***Analyst Coverage*** is the quarterly number of analysts covering the firm. ***Book/Market Ratio*** reports the yearly Book/Market of the firm. ***Retail Ownership*** is calculated as one minus institutional Ownership reported quarterly in form 13F of Thomson Reuters. ***Price Volatility*** is the standard deviation of stock price. ***Diff1*** presents the difference of mean between small and medium trades, ***Diff2*** presents the difference of mean between medium and large trades, and ***Diff3*** presents the difference of mean between small and large trades. Panel B presents trading data for each trading group and is based on buy/sell trade classifications. Trades have been binned into three trading groups. Small trades include trades up to $10,000; Medium trades include trades between $10,000 and $50,000; and large trades include trades greater than $50,000. Trading data is reported for the normal trading hours (09:30 AM – 4 PM). All the trade groups have been adjusted for inflation using the consumer price index.

Panel A: Firm-level data for each trade size

Variables	**Small Trades**						***Diff1***
	Mean	**SD**	**P25**	**P50**	**P75**	**Max**	
NOF (000)	0.16	0.4	0.04	0.03	0.13	76.5	0.01***
Abnormal Turnover	1.01	0.65	0.61	0.86	1.21	4.12	0.04***
Firm Size (Billions)	7.59	27.84	0.33	1.18	4.09	648	-3.16***
Firm Age	22.88	17.67	9	19	31	67	-1.76***
Analyst Coverage	6.36	6.37	2	4	9	56	-1.80***
Book/Market Ratio	0.47	0.41	0.153	0.38	0.71	2	0.05***
Retail Ownership	0.31	0.28	0.08	0.22	0.46	0.09	0.07***
Price Volatility	0.19	0.21	0.07	0.13	0.23	1.23	-0.03***
	Medium Trades						***Diff2***
	Mean	**SD**	**P25**	**P50**	**P75**	**Max**	
NOF (000)	0.05	0.31	0	0.04	0.02	56.7	-0.01***
Abnormal Turnover	0.97	0.81	0.46	0.76	1.21	4.87	-0.01***
Firm Size (Billions)	10.75	32.99	0.86	2.39	7.18	648	-8.08***
Firm Age	24.64	18.73	10	20	35	67	-2.76***
Analyst Coverage	8.16	6.7	3	7	12	56	-2.58***
Book/Market Ratio	0.42	0.38	0.14	0.34	0.62	1.81	0.04***
Retail Ownership	0.24	0.24	0.06	0.16	0.32	0.96	0.02***
Price Volatility	0.22	0.26	0.05	0.14	0.28	1.53	-0.05***
	Large Trades						***Diff3***
	Mean	**SD**	**P25**	**P50**	**P75**	**Max**	
NOF (000)	0.06	0.68	-0.07	0.01	0.04	83.9	0.10***
Abnormal Turnover	0.98	1.13	0.42	0.64	1.05	7.57	0.03***
Firm Size (Billions)	18.83	44.38	1.95	5.48	15.87	648	-11.24***
Firm Age	27.4	19.81	12	22	43	67	-4.52***
Analyst Coverage	10.74	7.22	5	10	15	56	-4.38***
Book/Market Ratio	0.38	0.36	0.12	0.29	0.55	1.66	0.09***
Retail Ownership	0.22	0.23	0.06	0.15	0.28	0.96	0.09***
Price Volatility	0.27	0.38	0.01	0.14	0.35	2.22	-0.08***

Panel B: Trading Data by Trade Groups and Buy/Sell Classifications

Description	**Small Trades**		**Medium Trades**		**Large Trades**	
No. of Trades (Millions)	10,281.56		246.59		12.79	
Avg. Value (Thousand)	3.54		47.25		498.95	
Avg. Volume	116.94		2,152.37		14,492.14	
	Buy	**Sell**	**Buy**	**Sell**	**Buy**	**Sell**
No. of Trades (Millions)	6,359.16	3,920.93	156.04	89.84	8.07	4.15
Avg. Value (Thousand)	3.53	3.53	45.62	45.68	461.51	435.09
Avg. Volume	113.61	117.36	1,829.08	2,019.92	11,835.01	14,250.69

Table 2.3: Net Order Flow and Social Media Attention

Notes: The table reports the contemporaneous fixed effect regressions during the trading hours (09:30 AM – 4 PM ET). The dependent variable is net order flow, and the explanatory variable is social media attention (SMA) in all the regressions. Trades are classified into three groups. Small trades include trades up to $10,000; Medium trades include trades between $10,000 and $50,000; and large trades include trades greater than $50,000. All the trade groups have been adjusted for inflation on the basis of the 1991 consumer price index. Variable definitions are presented in Appendix 2.1. *, **, and *** represent significance at 10%, 5%, and 1%, respectively. Standard errors are robust to heteroscedasticity and clustered at the firm-level are reported in brackets.

	Small Trades			**Medium Trades**			**Large Trades**		
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***
SMA	0.074*** [0.0034]	0.074*** [0.0034]	0.071*** [0.0033]	0.023*** [0.0022]	0.023*** [0.0022]	0.022*** [0.0020]	-0.001 [0.0013]	-0.001 [0.0013]	-0.001 [0.0013]
Abnormal Turnover	0.208*** [0.0105]	0.209*** [0.0105]	0.192*** [0.0098]	0.076*** [0.0055]	0.076*** [0.0055]	0.066*** [0.0052]	0.078*** [0.0066]	0.078*** [0.0066]	0.078*** [0.0067]
NOF_{t-1}	0.387*** [0.0325]	0.387*** [0.0325]	0.382*** [0.0321]	0.483*** [0.0561]	0.483*** [0.0560]	0.477*** [0.0554]	-0.050 [0.0497]	-0.050 [0.0497]	-0.050 [0.0497]
Abs. Abnormal Return	0.040*** [0.0032]	0.040*** [0.0032]	0.007* [0.0038]	0.017*** [0.0026]	0.017*** [0.0027]	-0.002 [0.0039]	0.002 [0.0029]	0.002 [0.0028]	0.001 [0.0032]
Avg. $Returns_{250}$	0.018*** [0.0064]	0.018*** [0.0063]	0.014** [0.0067]	0.006** [0.0028]	0.006** [0.0028]	0.005* [0.0028]	-0.000 [0.0046]	-0.000 [0.0045]	-0.001 [0.0047]
Firm Age		-0.006*** [0.0023]	-0.008*** [0.0023]		0.002 [0.0020]	0.000 [0.0016]		-0.000 [0.0014]	0.001 [0.0011]
Analyst Coverage		-0.013*** [0.0026]	-0.012*** [0.0024]		-0.001 [0.0016]	0.000 [0.0019]		0.004 [0.0028]	0.004 [0.0026]
Firm Size		-0.002 [0.0030]	-0.024*** [0.0035]		-0.007 [0.0067]	-0.016** [0.0078]		-0.004 [0.0028]	-0.004 [0.0029]
Book/Market Ratio			0.005 [0.0030]			0.001 [0.0016]			-0.005*** [0.0015]
Retail Ownership			0.017*** [0.0044]			0.005 [0.0039]			0.000 [0.0039]
Price Volatility			0.545*** [0.0525]			0.236*** [0.0607]			0.012 [0.0076]
Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Adj R-squared	0.2416	0.2417	0.2515	0.2715	0.2716	0.2790	0.0226	0.0226	0.0227
No. of Firms	2675	2675	2675	2675	2675	2675	2635	2635	2635
Observations	3,080,788	3,080,788	3,080,788	2,129,766	2,129,766	2,129,766	1,096,314	1,096,314	1,096,314

Table 2.4: Next Day Net Order Flow and Social Media Attention

Notes: The table reports the fixed effect regression results of next-day net order flow and social media attention during the trading hours (09:30 AM – 04 PM ET). The dependent variable is net order flow, and the explanatory variable is social media attention (SMA) in all the regressions. Trades are classified into three groups. Small trades include trades up to $10,000; Medium trades include trades between $10,000 and $50,000; and large trades include trades greater than $50,000. All the trade groups have been adjusted for inflation on the basis of the 1991 consumer price index. t+1, t+2, and t+5 represent the net order flow 1, 2, and 5 days ahead. Variable definitions are presented in Appendix 2.1. *, **, and *** represent significance at 10%, 5%, and 1%, respectively. Standard errors are robust to heteroscedasticity and clustered at the firm-level are reported in brackets.

	Small Trades			**Medium Trades**			**Large Trades**		
	t+1	***t+2***	***t+5***	***t+1***	***t+2***	***t+5***	***t+1***	***t+2***	***t+5***
SMA	0.003* [0.0018]	-0.007*** [0.0010]	-0.009*** [0.0007]	0.000 [0.0007]	-0.003*** [0.0007]	-0.003*** [0.0005]	0.002** [0.0006]	-0.000 [0.0005]	0.001*** [0.0004]
Abnormal Turnover	0.041*** [0.0102]	0.034*** [0.0055]	0.024*** [0.0039]	-0.003 [0.0062]	-0.006 [0.0071]	-0.008 [0.0065]	0.012*** [0.0035]	0.006*** [0.0012]	-0.008*** [0.0015]
NOF_{t-1}	0.414*** [0.0374]	0.290*** [0.0197]	0.178*** [0.0146]	0.494*** [0.0569]	0.396*** [0.0647]	0.286*** [0.0567]	-0.054 [0.0538]	-0.010 [0.0188]	0.009 [0.0119]
Abs. Abnormal Return	-0.015*** [0.0011]	-0.020*** [0.0012]	-0.018*** [0.0012]	-0.010*** [0.0020]	-0.010*** [0.0018]	-0.010*** [0.0018]	-0.001 [0.0016]	-0.001 [0.0014]	-0.001 [0.0011]
Avg. $Returns_{250}$	0.003 [0.0037]	0.007* [0.0036]	0.010*** [0.0029]	0.001 [0.0014]	0.002 [0.0015]	0.002** [0.0012]	-0.002 [0.0034]	-0.002 [0.0026]	-0.001 [0.0015]
Firm Age	-0.004 [0.0023]	-0.005* [0.0028]	-0.006* [0.0033]	-0.001 [0.0016]	-0.001 [0.0019]	-0.002 [0.0022]	0.000 [0.0014]	0.000 [0.0012]	-0.000 [0.0011]
Analyst Coverage	-0.009*** [0.0027]	-0.011*** [0.0027]	-0.011*** [0.0028]	0.000 [0.0019]	0.001 [0.0023]	0.001 [0.0028]	0.007 [0.0043]	0.006* [0.0032]	0.005** [0.0024]
Firm Size	0.003 [0.0031]	0.006 [0.0039]	0.008* [0.0045]	-0.009 [0.0066]	-0.010 [0.0077]	-0.011 [0.0089]	-0.008* [0.0045]	-0.008** [0.0035]	-0.006** [0.0027]
Book/Market Ratio	-0.002 [0.0020]	-0.003 [0.0024]	-0.003 [0.0026]	0.001 [0.0011]	0.001 [0.0013]	0.000 [0.0016]	-0.004*** [0.0015]	-0.004*** [0.0013]	-0.003*** [0.0013]
Retail Ownership	-0.014*** [0.0034]	-0.019*** [0.0042]	-0.022*** [0.0048]	-0.006* [0.0034]	-0.008* [0.0041]	-0.010* [0.0055]	0.003 [0.0056]	0.002 [0.0045]	0.001 [0.0036]
Price Volatility	0.087*** [0.0244]	0.040*** [0.0154]	0.032** [0.0141]	0.047** [0.0194]	0.018 [0.0146]	0.015 [0.0125]	0.025*** [0.0085]	0.024*** [0.0084]	0.019*** [0.0074]
Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Adj. R-squared	0.1942	0.0995	0.0446	0.2618	0.1738	0.1005	0.0191	0.0168	0.0160
No. of Firms	2675	2675	2675	2675	2675	2675	2635	2635	2635
Observations	3,080,788	3,080,788	3,080,788	2,129,766	2,129,766	2,129,766	1,096,314	1,096,314	1,096,314

Table 2.5: Net Order Flow, Social Media Attention and News Coverage

Notes: The table reports the moderating effect of media coverage significant news developments on the net order flow (NOF) during the trading hours (09:30 AM – 4 PM ET). The dependent variable is NOF, and the explanatory variable is social media attention (SMA) in all the regressions. ***has_media*** is the dummy variable equals one if the firm-specific tweets contain video/images in the text, otherwise 0. ***Sig. Developments*** are collected from Thomson Reuters news analysis which covers market-moving company events on a near real-time basis. ***News Coverage D.J.*** is the daily Dow Jones news stories count covering the sample firms. Trades are classified into three groups. Small trades include trades up to $10,000; Medium trades include trades between $10,000 and $50,000; and large trades include trades greater than $50,000. All the trade groups have been adjusted for inflation based on the 1991 consumer price index. Variable definitions are presented in Appendix 2.1. *, **, and *** represent significance at 10%, 5%, and 1%, respectively. Standard errors are robust to heteroscedasticity and clustered at the firm-level are reported in brackets.

	Small Trades			**Medium Trades**			**Large Trades**		
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***
SMA	0.064***	0.058***	0.058***	0.020***	0.018***	0.018***	-0.002	-0.001	-0.002
	[0.0031]	[0.0032]	[0.0030]	[0.0021]	[0.0020]	[0.0018]	[0.0015]	[0.0010]	[0.0014]
has_media	0.272***	0.268***	0.261***	0.122***	0.121***	0.118***	0.023**	0.023**	0.026**
	[0.0294]	[0.0293]	[0.0295]	[0.0216]	[0.0216]	[0.0214]	[0.0105]	[0.0107]	[0.0111]
Sig. Developments	0.167***	-0.040	0.080**	0.041**	-0.023	0.004	0.018	0.038	0.018
	[0.0359]	[0.0413]	[0.0388]	[0.0180]	[0.0212]	[0.0195]	[0.0355]	[0.0526]	[0.0398]
News Coverage DJ	0.041***	0.039***	0.015***	0.010***	0.010***	-0.003	0.003	0.003	0.003
	[0.0042]	[0.0042]	[0.0031]	[0.0034]	[0.0034]	[0.0043]	[0.0025]	[0.0025]	[0.0024]
*SMA * Sig. Developments*		0.295***			0.088***			-0.027	
		[0.0292]			[0.0143]			[0.0238]	
*SMA * News Coverage D.J.*			0.061***			0.027***			-0.001
			[0.0052]			[0.0042]			[0.0033]
Abnormal Turnover	0.187***	0.185***	0.182***	0.065***	0.064***	0.063***	0.077***	0.077***	0.078***
	[0.0096]	[0.0095]	[0.0092]	[0.0051]	[0.0051]	[0.0051]	[0.0067]	[0.0067]	[0.0068]
NOF_{t-1}	0.380***	0.381***	0.382***	0.476***	0.477***	0.478***	-0.050	-0.050	-0.052
	[0.0319]	[0.0319]	[0.0321]	[0.0554]	[0.0554]	[0.0552]	[0.0497]	[0.0497]	[0.0494]
Abs. Abnormal Return	0.005	0.005	0.001	-0.003	-0.004	-0.004	0.000	0.000	0.001
	[0.0038]	[0.0038]	[0.0039]	[0.0039]	[0.0039]	[0.0039]	[0.0030]	[0.0032]	[0.0034]
$Avg.\ Returns_{250}$	0.013**	0.013**	0.013*	0.005*	0.005*	0.003	-0.001	-0.001	-0.002
	[0.0067]	[0.0067]	[0.0074]	[0.0028]	[0.0028]	[0.0028]	[0.0047]	[0.0046]	[0.0046]
Firm Age	-0.007***	-0.006***	-0.007***	0.001	0.001	0.000	0.001	0.001	0.001
	[0.0024]	[0.0024]	[0.0024]	[0.0016]	[0.0016]	[0.0016]	[0.0010]	[0.0011]	[0.0013]
Analyst Coverage	-0.013***	-0.013***	-0.013***	-0.000	-0.000	-0.002	0.004	0.004	0.002
	[0.0024]	[0.0024]	[0.0021]	[0.0018]	[0.0018]	[0.0018]	[0.0026]	[0.0026]	[0.0021]
Firm Size	-0.024***	-0.024***	-0.021***	-0.016**	-0.016**	-0.015*	-0.004	-0.004	-0.002
	[0.0035]	[0.0035]	[0.0034]	[0.0079]	[0.0079]	[0.0077]	[0.0029]	[0.0029]	[0.0026]
Book/Market Ratio	0.003	0.003	0.010***	0.001	0.001	0.004*	-0.006***	-0.006***	-0.005***
	[0.0030]	[0.0030]	[0.0027]	[0.0016]	[0.0016]	[0.0023]	[0.0015]	[0.0015]	[0.0014]
Retail Ownership	0.018***	0.018***	0.018***	0.005	0.005	0.002	0.000	0.000	-0.003
	[0.0044]	[0.0044]	[0.0039]	[0.0040]	[0.0040]	[0.0034]	[0.0039]	[0.0039]	[0.0036]
Price Volatility	0.536***	0.533***	0.525***	0.234***	0.233***	0.235***	0.011	0.011	0.012
	[0.0521]	[0.0518]	[0.0522]	[0.0605]	[0.0605]	[0.0612]	[0.0076]	[0.0076]	[0.0074]
Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Adj R-squared	0.2544	0.2557	0.2613	0.2798	0.2802	0.2829	0.0227	0.0228	0.0247
No. of Firms	2675	2675	2675	2675	2675	2675	2612	2612	2612
Observations	3,080,788	3,080,788	3,080,788	2,129,766	2,129,766	2,129,766	1,096,314	1,096,314	1,096,314

Table 2.6: Social Media Attention and Price Pressure

Notes: The table reports the results from Fama-Mcbeth's (1973) cross-sectional regressions between abnormal returns and social media attention (SMA). The dependent variable is abnormal return, calculated as 5x5 firm size and book to market ratio as excess return from the benchmark return reported in basis points. Variable definitions are presented in Appendix 2.1. *, **, and *** represent significance at 10%, 5%, and 1%, respectively. Standard errors are computed using the Newey-West formula with four lags and are reported in brackets.

	1	*2*	*3*	*4*	*5*
SMA	5.170*** [0.1346]	5.024*** [0.1300]	11.214*** [0.3822]	11.381*** [0.3840]	9.802*** [0.3463]
ASVI	0.360*** [0.0664]	0.255** [0.1037]	0.254** [0.1039]	0.253** [0.1047]	0.254** [0.1049]
Firm Size	0.522*** [0.0797]	-0.035 [0.0727]	-0.044 [0.0735]	-0.041 [0.0735]	-0.043 [0.0733]
BSI_{Retail}		16.218*** [0.1852]	16.209*** [0.1851]	16.266*** [0.1866]	16.253*** [0.1865]
Retail Ownership		0.314** [0.1426]	0.319** [0.1426]	0.310** [0.1424]	0.306** [0.1423]
*SMA * Firm Size*			-0.854*** [0.0429]	-0.873*** [0.0427]	-0.630*** [0.0372]
SMA * BSI_{Retail}				2.045*** [0.1052]	2.095*** [0.1060]
*SMA * Retail Ownership*					1.822*** [0.0887]
Firm Age	0.874*** [0.1038]	1.220*** [0.0962]	1.212*** [0.0962]	1.210*** [0.0962]	1.207*** [0.0962]
Book/Market Ratio	0.497*** [0.1747]	0.781*** [0.1699]	0.779*** [0.1699]	0.773*** [0.1700]	0.772*** [0.1701]
Analyst Coverage	-0.518*** [0.1577]	-0.392** [0.1524]	-0.386** [0.1524]	-0.383** [0.1522]	-0.384** [0.1522]
Adv/Sales Ratio	0.151 [0.3169]	0.064 [0.3129]	0.061 [0.3119]	0.062 [0.3119]	0.070 [0.3118]
Avg. $Returns_{250}$	8.014*** [0.2835]	8.063*** [0.2764]	8.060*** [0.2763]	8.050*** [0.2760]	8.048*** [0.2759]
Abs. Abnormal Return	7.695*** [1.1180]	7.611*** [1.0956]	7.553*** [1.0955]	7.499*** [1.0959]	7.444*** [1.0954]
No. of Firms	2675	2675	2675	2675	2675
Observations	3,080,788	3,080,788	3,080,788	3,080,788	3,080,788

Table 2.7: Net Order Flow and Tweet Reach

Notes: The table reports the moderating effect of Tweets Reach on net order flow during the trading hours (09:30 AM – 4 PM ET). ***Tweets Reach*** is measured as the sum of distinct users' followers who posted firm-specific tweets. The dependent variable is net order flow, and the explanatory variable is social media attention (SMA) in all the regressions. Trades are classified into three groups. Small trades include trades up to $10,000; Medium trades include trades between $10,000 and $50,000; and large trades include trades greater than $50,000. All the trade groups have been adjusted for inflation based on the 1991 consumer price index. The dependent variable is net order flow in all the regressions. Variable definitions are presented in Appendix 2.1. *, **, and *** represent significance at 10%, 5%, and 1%, respectively. Standard errors are robust to heteroscedasticity and clustered by the firms are reported in brackets.

Terciles	***Panel A: Small Trades***							
	SMA	**Sig. Development**	**News Coverage DJ**	**Firm Level Controls**	**Stock Level Controls**	**Adj. R-squared**	**No. of Firms**	**Observations**
1	0.034*** [0.0021]	0.085*** [0.0185]	0.012 [0.0102]	Yes	Yes	0.2922	2675	661,165
2	0.050*** [0.0028]	0.035* [0.0189]	0.029*** [0.0093]	Yes	Yes	0.2726	2675	643,214
3	0.162*** [0.0110]	0.182*** [0.0651]	0.273*** [0.0269]	Yes	Yes	0.2367	2675	646,218
	Panel B: Medium Trades							
	SMA	**Sig. Development**	**News Coverage DJ**	**Firm Level Controls**	**Stock Level Controls**	**Adj. R-squared**	**No. of Firms**	**Observations**
1	0.013*** [0.0020]	0.034** [0.0152]	-0.011 [0.0099]	Yes	Yes	0.3828	2671	523,038
2	0.021*** [0.0031]	-0.006 [0.0136]	0.004 [0.0080]	Yes	Yes	0.3539	2668	511,586
3	0.054*** [0.0051]	0.040 [0.0317]	0.068*** [0.0170]	Yes	Yes	0.2520	2663	514,180
	Panel C: Large Trades							
	SMA	**Sig. Development**	**News Coverage DJ**	**Firm Level Controls**	**Stock Level Controls**	**Adj. R-squared**	**No. of Firms**	**Observations**
1	0.000 [0.0007]	-0.012 [0.0318]	-0.002 [0.0080]	Yes	Yes	0.0066	2485	308,626
2	0.000 [0.0007]	-0.003 [0.0176]	0.016 [0.0141]	Yes	Yes	0.0094	2453	304,323
3	-0.000 [0.0046]	0.047 [0.0649]	0.020** [0.0100]	Yes	Yes	0.0228	2439	304,834

Table 2.8: Net Order Flow and Tweets Concentration

Notes: The table reports the moderating effect of Tweets Concentration on net order flow during the trading hours (09:30 AM – 4 PM ET). ***Tweets Concentration*** is defined as the total number of unique users posting firm-specific tweets. Trades are classified into three groups. Small trades include trades up to $10,000; Medium trades include trades between $10,000 and $50,000; and large trades include trades greater than $50,000. All the trade groups have been adjusted for inflation based on the 1991 consumer price index. The dependent variable is net order flow, and the explanatory variable is social media attention (SMA) in all the regressions. Twitter activity for each stock is sorted into terciles based on Tweets Concentration. Tercile 1 – 3 represents tweets concentration from the lowest to the highest level, respectively. Variable definitions are presented in Appendix 2.1. *, **, and *** represent significance at 10%, 5%, and 1%, respectively. Standard errors are robust to heteroscedasticity and clustered by the firms are reported in brackets.

Terciles	*Small Trades*							
	SMA	**Sig. Development**	**News Coverage DJ**	**Firm Level Controls**	**Stock Level Controls**	**Adj. R-squared**	**No. of Firms**	**Observations**
1	0.023*** [0.0016]	0.069*** [0.0178]	0.015 [0.0092]	Yes	Yes	0.3000	2675	876,539
2	0.053*** [0.0032]	-0.008 [0.0224]	0.047*** [0.0113]	Yes	Yes	0.2561	2601	534,688
3	0.167*** [0.0133]	0.140** [0.0676]	0.331*** [0.0266]	Yes	Yes	0.2554	2675	539,370
	Medium Trades							
	SMA	**Sig. Development**	**News Coverage DJ**	**Firm Level Controls**	**Stock Level Controls**	**Adj. R-squared**	**No. of Firms**	**Observations**
1	0.011*** [0.0023]	0.017* [0.0102]	-0.003 [0.0068]	Yes	Yes	0.4396	2674	662,748
2	0.021*** [0.0036]	-0.005 [0.0157]	0.002 [0.0176]	Yes	Yes	0.3418	2589	444,210
3	0.056*** [0.0055]	0.020 [0.0355]	0.083*** [0.0214]	Yes	Yes	0.2740	2650	441,839
	Large Trades							
	SMA	**Sig. Development**	**News Coverage DJ**	**Firm Level Controls**	**Stock Level Controls**	**Adj. R-squared**	**No. of Firms**	**Observations**
1	0.000 [0.0006]	-0.004 [0.0310]	-0.005 [0.0059]	Yes	Yes	0.0037	2518	375,703
2	0.000 [0.0012]	-0.007 [0.0130]	0.007 [0.0108]	Yes	Yes	0.0084	2379	270,538
3	-0.003 [0.0051]	0.036 [0.0637]	0.033*** [0.0103]	Yes	Yes	0.0301	2396	271,581

Table 2.9: Tweets Characteristics and Net Order Flow

Notes: The table reports the regression results of tweets characteristics and net order flow during the trading hours (09:30 AM – 4 PM ET). ***Verified*** is the number of verified users who post firm-specific tweets, and ***Non-Verified*** is the number of non-verified users who post firm-specific tweets. ***Replies > 0*** counts the number of firm-specific tweets with at least one reply compared to ***Replies = 0*** with no replies. ***Retweets > 0*** counts the number of firm-specific tweets retweeted at least once compared to ***Retweets = 0*** with no retweets. The dependent variable is net order flow, and the explanatory variable is social media attention (SMA) in all the regressions. Trades are classified into three groups. Small trades include trades up to $10,000; Medium trades include trades between $10,000 and $50,000; and large trades include trades greater than $50,000. All the trade groups have been adjusted for inflation based on the 1991 consumer price index. ***Panel A*** presents SMA and net order flow for all the tweets with cashtags posted by verified and non-verified users, respectively. ***Panel B*** presents SMA and net order flow for all the tweets with cashtags with at least one reply and no reply, respectively, and ***Panel C***, presents SMA and net order flow for all the tweets with cashtags that have been retweeted at least once and not retweeted at all. Variable definitions are presented in Appendix 2.1. *, **, and *** represent significance at 10%, 5%, and 1%, respectively. Standard errors are robust to heteroscedasticity and clustered by the firms are reported in brackets.

Panel A	**Verified**			**Non-Verified**		
	Small Trades	***Medium Trades***	***Large Trades***	***Small Trades***	***Medium Trades***	***Large Trades***
SMA	0.429***	0.139***	0.001	0.032***	0.007***	0.001***
	[0.0257]	[0.0156]	[0.0094]	[0.0009]	[0.0004]	[0.0002]
ASVI	0.066***	0.026***	0.013***	0.002***	0.001***	0.000
	[0.0104]	[0.0060]	[0.0050]	[0.0002]	[0.0001]	[0.0002]
Sig. Developments	0.359***	0.092*	0.064	0.087***	0.015***	0.008
	[0.1207]	[0.0551]	[0.0899]	[0.0075]	[0.0035]	[0.0070]
News Coverage DJ	0.271***	0.045*	0.032**	0.035***	0.007***	0.005***
	[0.0457]	[0.0267]	[0.0133]	[0.0029]	[0.0015]	[0.0013]
Panel B	**Replies > 0**			**Replies = 0**		
	Small Trades	***Medium Trades***	***Large Trades***	***Small Trades***	***Medium Trades***	***Large Trades***
SMA	0.336***	0.110***	0.002	0.030***	0.006***	0.001***
	[0.0197]	[0.0119]	[0.0072]	[0.0008]	[0.0003]	[0.0002]
ASVI	0.044***	0.019***	0.011***	0.002***	0.001***	-0.000
	[0.0068]	[0.0042]	[0.0039]	[0.0002]	[0.0001]	[0.0002]
Sig. Developments	0.450***	0.131**	0.075	0.098***	0.013***	0.001
	[0.1133]	[0.0520]	[0.0863]	[0.0077]	[0.0025]	[0.0043]
News Coverage DJ	0.313***	0.061**	0.032**	0.046***	0.011***	0.006***
	[0.0430]	[0.0256]	[0.0127]	[0.0030]	[0.0013]	[0.0013]
Panel C	**Retweets > 0**			**Retweets = 0**		
	Small Trades	***Medium Trades***	***Large Trades***	***Small Trades***	***Medium Trades***	***Large Trades***
SMA	0.268***	0.087***	0.002	0.023***	0.004***	0.001***
	[0.0151]	[0.0089]	[0.0051]	[0.0007]	[0.0002]	[0.0002]
ASVI	0.036***	0.015***	0.007***	0.001***	0.000***	0.000
	[0.0046]	[0.0030]	[0.0027]	[0.0002]	[0.0001]	[0.0002]
Sig. Developments	0.338***	0.092**	0.049	0.072***	0.009***	0.007***
	[0.0816]	[0.0381]	[0.0640]	[0.0061]	[0.0022]	[0.0023]
News Coverage DJ	0.272***	0.058***	0.030***	0.035***	0.009***	0.003***
	[0.0332]	[0.0205]	[0.0101]	[0.0025]	[0.0009]	[0.0010]

Table 2.10: Social media attention and alternative proxies of attention

Notes: The table reports the regression results of Social Media Attention (SMA) and alternative proxies of attention. The dependent variable is SMA during the trading hours (09:30 AM – 4 PM ET). ***Firm Size*** the natural log of the market value of the firm's equity. ***Analyst Coverage*** is the number of analysts following the firm. ***Sig. Developments*** are collected from Thomson Reuters news analysis which covers market-moving company events on a near real-time basis. ***News Coverage DJ*** is the daily Dow Jones news stories count covering the sample firms. ***ASVI*** is the natural log of one plus the change in median search volume index from Google Analytics. ***AIA*** abnormal institutional attention is downloaded from Bloomberg Terminal. Variable definitions are presented in Appendix 2.1. *, **, and *** represent significance at 10%, 5%, and 1%, respectively. Standard errors are robust to heteroscedasticity and clustered by the firms are reported in brackets.

	1	***2***	***3***	***4***	***5***	***6***
Firm Size	0.001 [0.0010]	0.001 [0.0010]	0.001 [0.0010]	0.001 [0.0011]	0.019*** [0.0016]	0.019*** [0.0016]
Analyst Coverage		-0.005*** [0.0007]	-0.006*** [0.0007]	-0.008*** [0.0010]	-0.011*** [0.0013]	-0.011*** [0.0013]
Sig. Developments			0.659*** [0.0116]			
News Coverage DJ				0.113*** [0.0027]		
ASVI					-0.000 [0.0003]	
AIA						0.367*** [0.0065]
Abs. Abnormal Return	0.113*** [0.0015]	0.113*** [0.0015]	0.111*** [0.0015]	0.109*** [0.0014]	0.102*** [0.0013]	0.102*** [0.0013]
Abnormal Turnover	0.066*** [0.0012]	0.066*** [0.0012]	0.060*** [0.0011]	0.056*** [0.0011]	0.050*** [0.0009]	0.050*** [0.0009]
Intercept	-0.004 [0.0073]	-0.008 [0.0071]	-0.021*** [0.0072]	-0.006 [0.0080]	-0.200*** [0.0112]	-0.200*** [0.0112]
Fixed Effects	Y	Y	Y	Y	Y	Y
Adj. R-squared	0.0192	0.0192	0.0264	0.0299	0.0304	0.0304
No. of Firms	2675	2675	2675	2675	2675	2675
Observations	3,080,788	3,080,788	3,080,788	3,080,788	3,080,788	3,080,788

Table 2.11: Social Media Attention during Non-Trading Hours

Notes: The table reports regression results of social media attention (SMA) and net order flow during non-trading hours (4 PM – 09:30 AM ET). Non-trading hours further divided into pre-market period (12 AM – 09:29:59 AM ET) and post-market period (4 PM – 11:59:59 PM ET). The dependent variable is net order flow (NOF) during the trading hours. For the post-market period, the dependent variable is the next day NOF. However, for the pre-market period, the explanatory variable is preceding SMA within 24 hours period. Trades are classified into three groups. Small trades include trades up to $10,000; Medium trades include trades between $10,000 and $50,000; and large trades include trades greater than $50,000. All the trade groups have been adjusted for inflation based on the 1991 consumer price index. Variable definitions are presented in Appendix 2.1. *, **, and *** represent significance at 10%, 5%, and 1%, respectively. Standard errors are robust to heteroscedasticity and clustered by the firms are reported in brackets.

	Small Trades		***Medium Trades***		***Large Trades***	
	Pre-Market	***Post-Market***	***Pre-Market***	***Post-Market***	***Pre-Market***	***Post-Market***
SMA	0.044*** [0.0019]	0.048*** [0.0027]	0.013*** [0.0013]	0.015*** [0.0014]	-0.002** [0.0009]	-0.001 [0.0011]
Abnormal Turnover	0.195*** [0.0097]	0.197*** [0.0098]	0.074*** [0.0055]	0.074*** [0.0055]	0.085*** [0.0074]	0.085*** [0.0074]
NOF_{t-1}	0.377*** [0.0315]	0.379*** [0.0318]	0.474*** [0.0556]	0.475*** [0.0556]	-0.050 [0.0497]	-0.050 [0.0497]
Abs. Abnormal Return	0.012*** [0.0040]	0.011*** [0.0039]	-0.002 [0.0046]	-0.002 [0.0045]	-0.001 [0.0037]	-0.001 [0.0038]
Avg. Returns$_{250}$	0.013** [0.0067]	0.013** [0.0067]	0.006* [0.0034]	0.006* [0.0034]	-0.001 [0.0057]	-0.001 [0.0057]
Firm Age	-0.008*** [0.0024]	-0.008*** [0.0024]	0.000 [0.0017]	0.000 [0.0017]	0.000 [0.0013]	0.000 [0.0013]
Analyst Coverage	-0.012*** [0.0024]	-0.011*** [0.0024]	0.001 [0.0021]	0.001 [0.0021]	0.005 [0.0031]	0.005 [0.0031]
Firm Size	-0.025*** [0.0036]	-0.025*** [0.0036]	-0.016** [0.0081]	-0.016** [0.0081]	-0.004 [0.0033]	-0.004 [0.0033]
Book/Market Ratio	0.005 [0.0030]	0.005 [0.0030]	0.001 [0.0017]	0.001 [0.0017]	-0.007*** [0.0018]	-0.007*** [0.0019]
Retail Ownership	0.018*** [0.0045]	0.018*** [0.0045]	0.003 [0.0042]	0.003 [0.0042]	-0.000 [0.0048]	-0.000 [0.0048]
Price Volatility	0.565*** [0.0529]	0.564*** [0.0529]	0.253*** [0.0630]	0.253*** [0.0630]	0.018** [0.0084]	0.018** [0.0084]
Fixed Effects	Y	Y	Y	Y	Y	Y
Adj. R-squared	0.2487	0.2490	0.2799	0.2800	0.0230	0.0230
Firms	2675	2675	2675	2675	2527	2527
Observations	3,079,861	3,079,861	1,968,694	1,968,694	947,617	947,617

Table 2.12: Ranking Analysis of Firm-Specific Tweets

Notes: The table presents the ranking analysis based on the firm-specific tweets for the sample firms. Panel A presents 50 most discussed U.S companies based on the firm's cashtags appearing in the tweets. Panel B shows the top 15 mentions in the firm-specific tweets of the sample firms. Panel C presents the top 15 Twitter trends based on Hashtags in the firm-specific tweets of the sample firms.

Panel A: 50 most discussed U.S companies on Twitter

Company	Tweets	Daily Avg	Company	Tweets	Daily Avg
Apple Inc	1,313,058	1700.85	Walmart Inc	159,532	126.81
Facebook Inc	993,253	854.04	Cisco Systems Inc	149,951	119.20
Twitter Inc	685,851	864.88	Agilent Technologies Inc	142,061	112.93
Tesla Inc	506,896	402.94	Starbucks Corp	140,173	111.43
Amazon.com Inc	477,943	482.77	Wells Fargo & Co	139,727	111.07
Netflix Inc	419,051	356.94	Hyatt Hotels Corp	138,402	110.02
Microsoft Corp	385,575	306.50	General Motors Co	131,512	104.54
Alphabet Inc	343,574	445.04	McDonald's Corp	130,547	103.77
Bank of America Corp	260,283	206.90	Verizon Communications	129,060	102.59
JPMorgan Chase & Co	222,801	177.11	Johnson & Johnson	126,456	100.52
Goldman Sachs Group Inc	220,844	175.55	Pfizer Inc	120,737	95.98
Intel Corp	214,536	170.54	Chipotle Mexican Grill Inc	120,622	140.75
International Business Machines Corp	214,505	170.51	eBay Inc.	118,731	94.38
Altaba Inc	201,515	160.19	Caterpillar Inc	118,450	94.16
GoPro Inc	194,157	248.60	J.C. Penney Company Inc.	117,037	93.03
Citigroup Inc	189,262	150.45	Celgene Corp	116,205	92.37
Gilead Sciences Inc	189,070	150.29	Micron Technology Inc.	115,866	92.10
LinkedIn Corp	186,043	149.79	Alcoa Corp	113,847	90.50
Visa Inc	185,775	147.67	Sarepta Therapeutics Inc	111,943	99.50
Walt Disney Co (The)	178,138	141.72	Chesapeake Energy Corp	111,198	88.39
Ford Motor Co	175,800	139.86	Nike Inc	109,614	87.13
Sprint Corp	172,886	137.54	Chevron Corp	108,821	86.50
AT&T Inc	169,207	134.50	Freeport-McMoRan Inc	107,024	85.14
General Electric Co	163,973	130.34	Coca-Cola Co (The)	106,535	84.69
Exxon Mobil Corp	160,733	127.77	salesforce.com Inc	106,378	84.56

Panel B: Top 15 Twitter Mentions | ***Panel C: Top 15 Twitter Hashtags***

Twitter Mentions	Counts	Daily Avg	Twitter Hashtags	Counts	Daily Avg
@jimcramer	102,868	56.30	#STOCKS	1,586,113	868.15
@CNBC	53,633	29.36	#STOCK	1,006,750	551.04
@simplestockqtes	42,291	23.15	#NASDAQ	575,488	314.99
@WSJ	41,314	22.61	#INVESTING	524,571	287.12
@kicksonfire	38,149	20.88	#STOCKMARKET	480,428	262.96
@CNBCFastMoney	32,204	17.63	#FINANCE	452,659	247.76
@TheStreet	31,180	17.07	#STOCKACTION	436,355	238.84
@YouTube	30,129	16.49	#PENNYSTOCKS	429,589	235.13
@YahooFinance	29,374	16.08	#INVEST	425,032	232.64
@Benzinga	26,603	14.56	#TRADEIDEAS	422,905	231.48
@SAI	25,778	14.11	#SHARE	410,806	224.85
@IBDinvestors	24,281	13.29	#APPLE	261,615	143.19
@RatingsNetwork	20,108	11.01	#TRADING	254,428	139.26
@jack	18,383	10.06	#SP500	98,152	53.72
@Reuters	16,405	8.98	#NYSE	89,798	49.15

Appendix 2.1: Variables Definitions

Variable	Source	Definitions
		Dependent Variables
NOF	DTAQ	For a given trade group, Net Order Flow is calculated using daily buy/sell trades. First, the sum of the market value of the buyer and seller-initiated trades are normalized by the lagged market capitalization of NASDAQ/NYSE. The net order flow is estimated as the difference between the normalized market value of buy and sell trades within each trade group.
Abnormal Returns	CRSP	Daily value of excess returns of Firm Size and Book-to-market value adjusted 5 x 5 portfolio returns.
		Trade Groups
Small		Trades up to \$10,000
Medium		Trades > \$10,000 and ≤ \$50,000
Large		Trades > \$50,000
		Social Media Variables
SMA	Twitter Streaming API	Social media attention is measured as the natural log of one plus the total number of firm-specific tweets at time *t*. Days with no Twitter activity have 0 value for tweets.
Tweet Reach		Sum of distinct users' followers who post firm-specific tweets.
Users' Concentration		The number of distinct users posting firm-specific tweets.
has_media		Dummy variable equals one if the tweets with cashtags also contain video/images in the text, otherwise 0.
User Experience		Numbers of years from the date of joining Twitter to the data of posting the first tweet with cashtag in our sample
Verified		Dummy variable equals one if the Twitter user has a verified profile, otherwise 0.
Retweet		Dummy variable equals one if the firm-specific tweets have been retweeted once otherwise 0.
Reply		Dummy variable equals one if the firm-specific tweets have been replied once otherwise 0.
Hashtags		The total number of unique hashtags posted by users for a specific firm per day.
Mentions		The total number of unique mentions posted by users for a specific firm per day.
		Media & Search Variables
Sig. Developments	Thomson Reuters Eikon	Thomson Reuters Significant Developments is a news analysis and filtering service that simplifies the news by providing concise summaries and categorizations of crucial, market-moving company events on a near real-time basis.

		We aggregate the number of significant developments posted by Thomson Reuters daily.
Variable	**Source**	**Definitions**
News Coverage D.J.	Lexis-Nexis	Firm-specific articles are issued by Dow Jones News Service and covered by the Lexis-Nexis database daily.
ASVI	Google Analytics	The abnormal search volume index is measured using Google search volume index (SVI) data. The data is aggregated at the firm-week level. ASVI is the natural log of one plus the change in median SVI.
AIA	Bloomberg	Abnormal Institutional Attention is downloaded from Bloomberg Terminal. Raw data is processed based on Ben-Rephael et al. (2017) with daily frequency.
	Baseline Controls	
Abs. Abnormal Returns		The absolute value of daily Abnormal Returns.
$Avg.Returns_{250}$		Average daily returns of the firm in the previous year.
Price Volatility	CRSP	Firm stock price standard deviation.
Abnormal Turnover		The daily stock trading volume ratio on the trading day to its average trading volume over the previous year.
BSI_{Retail}	DTAQ	Buy/Sell Imbalance is the difference between the volume of daily Buy and Sell trades.
Firm Age		Age of the firm since it was first added into the Compustat Database.
Firm Size		Log of the market of value of firm equity.
Book/Market Ratio	Compustat	Book to Market ratio of the firm.
$Dummy_{Ann}$		Dummy variable equals one on the day of quarterly earnings announcement otherwise 0.
Analyst Coverage	I/B/E/S	Total number of Analysts covering the firm in a year as per I/B/E/S database.
Retail Ownership	Thomson Reuters	The institutional ownership data is collected from Form 13F, which is available quarterly. Retail Ownership is measured as one minus institutional Ownership.

Chapter 3

3. Does Divergence of Opinion make better minds? Evidence from Social Media

*It is the mark of an educated person to search for the same kind of clarity in each topic to the extent that the nature of the matter accepts **it.** For it is similar to expect a mathematician to speak persuasively or for an orator to furnish clear proofs! Each person judges well what they know and is thus a good critic of those things, [to be a critic] one must be educated about everything.*

-- Aristotle, Nicomachean Ethics

3.1. Introduction

Disagreement among investors has emerged as the centerpiece of behavioral finance research in understanding its role to drive investors' trading in the financial markets (Hong & Stein, 1999; Kandel & Pearson, 1995; Karpoff, 1986; Varian, 1989). These studies provide corroborative evidence about the deviation from rational models and highlight the need to explore investor trading behaviors to unleash plausible evidence of variations from the rational models. However, there is mixed evidence about the role of investors' disagreement in financial markets. For example, Carlin et al. (2014) argue that greater disagreement is directly proportional to expected returns as investors could face higher uncertainty and adverse selection when a disagreement arises. This argument is consistent with Harris and Raviv (1993); Varian (1985), and Banerjee and Kremer (2010), who suggest that heterogeneous priors play an essential role in increasing disagreement, and consequently, require additional compensation and higher risk premiums. In contrast, Miller (1977) argues that the divergence of opinions in the market leads to lower risk premiums in the presence of short-sale constraints, and asset prices should reflect the valuation of optimists. Studies by Chen et al. (2002) and Yu (2011) found compelling evidence in favor of Miller's hypothesis. Either way, these studies highlight the critical role of disagreement in financial markets and how it affects asset prices.

Prior studies have mainly focused on investigating the predictive power of disagreement and its underlying behavioral explanations. However, little is known about the mechanisms by which disagreement among investors influences financial markets and whether disagreement among investors offers firm-specific information and contributes to the efficiency of capital allocation. Therefore, building on Hong and Stein (2007) theoretical framework and motivated by the role of disagreement in the financial markets, we explore whether disagreement among investors results in less return synchronicity, consequently providing more firm-specific information.

A critical challenge for researchers is to find a suitable proxy[29] for disagreement among investors. With the recent advances in technology, there has been a paradigm shift in how market participants consume information. Social media platforms[30] for investors have recently emerged as popular information-sharing platforms where investors can share ideas, learn investment techniques, and recommend stocks based on their analyses.[31] This increases investors' ability to communicate information. Investor-oriented social media platforms provide an opportunity to observe disagreement among investors, including the heterogeneity of investors and the salience of information signals. Compared to traditional proxies of disagreement such as abnormal trading volume, volatility, and analyst forecast dispersions, investor-oriented social media platforms can offer better insights for the following reasons. First, investor-oriented social media platforms can provide a unique opportunity to observe the heterogeneity[32] of investors. Second, such platforms can provide direct evidence of disagreement arising from investors' social interactions when disclosing their recommendations[33]. Third, the salience[34] of information signals is a decisive factor for investors to allocate their attention efficiently. On such platforms, the salience of information signals can also be observed directly. Following Cookson and Niessner (2019), we use a novel data set from one of the largest investor-oriented social media platforms, StockTwits, to construct a disagreement proxy that is entirely based on social interactions[35] among investors. StockTwits provides useful insights to observe social interactions among investors and the salience of their information signals.

---

[29] Veldkamp (2006) explores the role of information-driven comovement in financial markets and highlights the need for reliable proxies for information acquisition by investors.

[30] Social media platforms are an important byproduct of technological advances. For example, Facebook and Twitter revolutionized the concept of online interactions in the first decade of the 21st century. Similarly, StockTwits started its Twitter-like cashtags service in 2008 for investors, and later Twitter adopted the same technology by adding a cashtags service to its online platform in 2012.

[31] In addition to StockTwits, which is mainly used for brief discussions and ideas by anyone in the investment industry. SeekingAlpha is another prominent investor-oriented platform offering stock market analysis to its users since 2004.

[32] Heterogeneity of agents is a key ingredient of belief dispersion (Kandel & Pearson, 1995).

[33] Not all but StockTwits and Seeking Alpha are two investor-oriented platforms where investors can voluntarily disclose their recommendations.

[34] For details see Huang et al. (2018) and (Li et al., 2019).

[35] Hong and Stein (1999) highlight the need to develop an interaction-based model that incorporates the follow of information in the financial markets. Moreover, Hong et al. (2004) provide evidence that social interactions among investors increase stock market participation.

Return synchronicity has been widely used as a proxy for price informativeness.[36] Prior research argues that return synchronicity plays a pivotal role in understanding the extent to which a stock comoves with industry and market factors. Roll (1988) finds low R^2 statistics in the absence of any news, suggesting that the capitalization of private information through the trading activities of informed arbitrageurs. According to the price informativeness explanation, less return synchronicity reflects more firm-specific information. Therefore, investors will benefit when trading by using that firm-specific information to allocate their resources more efficiently (Durnev et al., 2003; Morck et al., 2000). Moreover, investors will benefit when they wish to diversify their risk and adjust their portfolio betas (Shiyang Huang et al., 2019).

Our results show that disagreement among investors on StockTwits decreases return synchronicity. Specifically, one standard deviation increase in disagreement results in a 5.7% decrease in return synchronicity, suggesting an inflow of firm-specific information into the financial markets. These results remain robust after controlling for the effect of media coverage, analyst coverage, and macroeconomic trends, in addition to other firm-level controls. Similarly, these results are robust after controlling for firms, time, and industry fixed effects. We also use two-dimensional clustering for firms and months to deal with any serial correlation at the firm level and any systematic shocks over time (Petersen, 2009). Our findings suggest that social interactions play a key role in influencing investors' behaviors in financial markets (Hirshleifer, 2019; Hong et al., 2004).

However, there is mixed evidence on the role of return synchronicity to predict stock price informativeness. For example, Chan et al. (2013) find that liquidity is positively associated with return synchronicity, suggesting low R^2 statistics reflect information asymmetry. Barberis et al. (2005) presented evidence against fundamentals-based views that suggest that the returns of the S&P 500 index comove more with the returns of the stocks already listed in the S&P 500 index, and the returns of the stocks not listed in the S&P 500 index comove more with the returns of the non-listed stocks.[37] They argue that irrational investors trade in financial markets based on category view, habitat view, and information-diffusion view. Therefore, frictions or noise could lead to comovement in the stock returns if

[36] For example, Sila et al. (2017) present that the reputation of independent directors is directly linked to increased price informativeness, Mathers et al. (2017) find firms with less synchronicity have better innovation outcomes.

[37] Green and Hwang (2009) studied comovement before and after stock splits and present evidence supporting Barberis et al. (2005). They argue that stocks move more with high-priced stocks before splits and low-priced stocks after splits, thus following a category view-based approach.

there are limits to arbitrage[38]. In addition, it is also possible that noise trading increases the magnitude of idiosyncratic pricing errors, which are responsible for low R^2 statistics. Therefore, noise can distort the measure of price informativeness and low R^2 statistics do not necessarily indicate deteriorating informational efficiency.

In our research setting, noise is of less concern since StockTwits is an investor-oriented platform. However, to further test whether disagreement on StockTwits provides firm-specific information, it is important to disentangle the difference between noise and price informativeness. To test this conjecture, we implement different tests. First, following Ayers and Freeman (2003), we combine disagreement with leads, contemporaneous, and lag changes in earnings to predict current stock returns. If disagreement among investors reflects noise, disagreement should not increase the price leads of earnings, and the impact should go in the opposite direction for post-earnings-announcement drift. In contrast, we find clear evidence that disagreement increases price leads of earnings which suggests that disagreement accelerates the pricing of future earnings and generates an improvement. Second, we explore individual investors' recommendation revisions on StockTwits. Hong and Stein (2007) provide evidence that suggests that disagreement initially arises due to heterogeneous priors and differential interpretations of information signals by investors in financial markets. Therefore, we assume that investors will revise their recommendations when they update their economic models,[39] based on either their priors or differential interpretation, increasing disagreement among investors and the inflows of firm-specific information in the financial markets. We find consistent results in support of this assumption.

Third, we test to what extent the impact of disagreement on return synchronicity varies across firms with different levels of media coverage.[40] Our findings highlight the two key roles of disagreement among investors on StockTwits. First, StockTwits acts as a catalyst by offering firm-specific information, when there is too little or no information available from traditional information channels, suggesting disagreement from investor-oriented platforms containing firm-specific information rather than market-wide information. These results are consistent with Roll (1988), who present evidence in favor of low R^2 in the absence of news. Second,

[38] Based on the frictiona-based and sentiment-based explanations of comovement, disagreement among investors may increase frictions or sentiment, thereby inducing a high R^2 statistics.

[39] This argument is also in line with Kandel and Pearson (1995) and Hong and Stein (1999) suggesting that investors revise their economic models when their marginal utility of consuming new information is higher than the already available information in the financial markets.

[40] Unlike Dang et al. (2020), in addition to using aggregate media coverage as a control variable, we segregate media coverage based on news types, news topics, and news sources.

StockTwits acts as an intermediary and amplifies the flows of firm-specific information into financial markets when there is greater media coverage. These results are consistent with Dang et al. (2020) and complement their findings by highlighting the role of social media platforms in the financial markets.

Endogeneity could be a concern in our results as there may be selective coverage of firms on StockTwits, which depends on several exogenous factors as well as biased recommendations due to affiliations with the sample firms. To deal with the endogeneity issue, we first use a two-stage least square (2SLS) instrumental variable approach. We then use two instrumental variables for disagreement. First, we construct a unique proxy of local investors, *Proximity*, which captures the social distances between investors and the firms' headquarters for whom they have been discussing and sharing ideas on StockTwits. Our second instrument is motivated by the role of labor unions in US firms, defined as the total number of disputes between labor unions and firms aggregated monthly.

These instruments provide an independent source of exogenous variations for each endogenous regressor and meet the criteria for a valid instrument. First, for the relevance restriction, local investor and labor issues can provide a higher level of firm-specific information. Since StockTwits is an investor-oriented platform, it gives local investors a unique opportunity to share their opinions. Similarly, *Labor_Issues* is positively associated with disagreement as it can exacerbate the number of discussions on social media platforms.

Second, for the exclusion restriction, it is doubtful that price-based comovement can directly affect investors' locations and labor unions. In other words, both instruments should affect only return synchronicity via social media channels as information from these variables is sublimed further by a large network of investors who are experts in their fields. Consequently, such social interactions exacerbate the flow of firm-specific information. Our results, based on the 2SLS approach, force the exogeneous portion of disagreement to explain return synchronicity and leave our main results unchanged. In addition, considering that disagreement on StockTwits is a choice for investors, this can depend on several exogenous factors.[41] Under such circumstances, self-selection bias could be an issue that influences OLS

[41] For example, investors' education, investment type, background, and willingness to participate in different types of communication.

estimates (Heckman, 1979). To address this concern, we implement the two-stage Heckman (1979) selection, model. Our results remain unchanged after controlling for self-selection bias.

Next, we investigate the role of disagreement in influencing the firm-information environment. The Information environment plays a vital role in determining the influence of agents' learning behaviors in financial markets (Vega, 2006). The intuition behind this idea is that if disagreement among investors on StockTwits exacerbate the flows of firm-specific information, they should assist stakeholders by offering firm-specific information to correctly calculate the fair value of the firm (Lin et al., 2011), leaving less room for managers to conceal self-serving behaviors (Jin & Myers, 2006). Therefore, we would expect the impact of disagreement on return synchronicity to be more pronounced when the firm-information environment is less transparent.

To test this hypothesis, we first use discretionary accruals as a proxy for firm opacity (Hutton et al., 2009). Second, we explore the extent to which the impact of disagreement on return synchronicity varies with firms' diversity. Bushman, Chen, et al. (2004) present evidence that the firm's diversity limits the transparency of firms' operations for outside investors. Therefore, we define diversity as the number of business segments and geographic locations the firm operates in and use it as a second proxy for the information environment (Markarian & Parbonetti, 2007). Third, we examine differences in the impact of disagreement on return synchronicity across firms with different Industry Concentration levels.[42] Ali et al. (2014) argue that firms in highly concentrated industries disclose less information since the proprietary cost of information is higher in those industries.

Finally, we explore firms that are subject to insider trading activity. Piotroski and Roulstone (2004) present evidence that insider trading activity is a vital private information source for investors. Therefore, we expect that disagreement can facilitate the incorporation of private information into financial markets. Overall, our results show that disagreement has a higher impact on return synchronicity for opaque firms, firms with greater diversity, firms with greater industry concentration, and firms with more insider trades. These findings provide substantial evidence that investors' social media platforms facilitate the incorporation of firm-specific information when firms have a less transparent information environment.

[42] We use total assets to calculate industrial concentration based on the Herfindahl–Hirschman index (HHI). However, our results remain consistent when we use total sales to calculate industry concentration.

One of the key characteristics of social media is its attention-grabbing features, known as salience (Fiske & Taylor, 1991). It is pertinent to note that the scarcity of cognitive resources limits investors' ability to allocate their attention (Kahneman, 1973). Therefore, salience plays a pivotal role in guiding investors to allocate their attention. Previous studies have mainly discussed the impact of limited attention without explaining the effects of salience (Antweiler & Frank, 2004; Dang et al., 2020; Ding et al., 2019). Therefore, our next strand of investigation is to understand the role of salience. Specifically, we aim to disentangle the impact of disagreement associated with salience. We classify the salience based on information signals and the heterogeneity of investors.

The first salience group, which is based on information signals, is further divided into two subgroups based on their network and social media attention (SMA), where the network is defined as the reach of information signals, and SMA is further divided into the number of ideas on StockTwits, the popularity of those ideas, and discussion threads created by investors on StockTwits. The results show that when investors with large numbers of followers post ideas, other investors follow their lead, thus increasing their influence. Consequently, the higher salience of information signals attracts more audiences to interact with each other, increasing the level of disagreement and prompting higher inflows of firm-specific information into financial markets.

The second salience group, based on investors' heterogeneity, is further divided into three subgroups based on unique investors' presence, investors' self-disclosed investment experience, and approaches. First, our findings suggest that unique investors' arrival increases the impact of disagreement on return synchronicity. These results are consistent with Kandel and Pearson (1995), who argue that heterogeneity of agents is a key ingredient for disagreement among the investors. Second, we calculate within-group disagreement among investors based on their investment experience (professional, intermediate, or novice). Our results show that although professional investors take a lead role in assisting investors by facilitating inflows of firm-specific information, intermediate and novice investors also play a pivotal role. Third, we calculate within-group disagreement among investors based on their investment approaches (momentum, technical, fundamental, or value). Our results suggest that momentum and technical investors provide higher inflows of firm-specific information than investors with fundamental or value investment approaches. These findings are consistent with previous literature highlighting the role of media in increasing momentum investing (Hillert et al., 2014).

Our research contributes to the existing literature by offering direct evidence that discussions on investors-oriented social media platforms such as StockTwits provide firm-specific information. Overall, this study contributes to the following areas. First, unlike previous studies that only focus on the role of social media in financial markets to predict volumes and returns, our study provides an essential piece of the puzzle by explaining that a social media platform for investors can predict financial markets because it allows firm-specific information to flow to investors who actively participate in discussions on such platforms and update their priors based on available information. We accomplish this by using disagreement as a unique proxy for discussions among investors on StockTwits and providing substantial evidence that such disagreement can predict stock-return synchronicity. These findings are consistent with Hong and Stein (2007) disagreement model. To the best of our knowledge, this is the first study to provide evidence that chatter on social media platforms assists investors by offering firm-specific information.

Second, this study contributes to the existing literature investigating the firm information environment's role in financial markets. These results provide substantial evidence by highlighting the role of disagreement among investors in a less transparent information environment, thus suggesting that for firms with higher informational opacity, greater diversity, higher industry concentration, and more insider trading, social media platforms for investors act as a catalyst to assist investors by providing firm-specific information.

Third, our study contributes to the emerging literature on the role of the salience of information signals in financial markets. We show that a large social network of investors and a large number of unique ideas posted on StockTwits amplify the impact of disagreement on return synchronicity. Moreover, dissecting investors' heterogeneity on StockTwits further suggests that although geographical background and investors' experience matter, what matters more is the quality of information signals arriving in financial markets and investors' heterogeneity (Kandel & Pearson, 1995).

Finally, our study contributes to distinguishing the debate on whether stock prices for companies with a low R^2 statistics is more informative. This study provides substantial evidence using a unique dataset of investor-oriented platforms that less synchronicity can reflect higher stock price informativeness.

Our findings are consistent with the existing literature in providing substantive evidence that social media platforms for investors can predict financial markets. However, no previous study has discussed whether disagreement on social media platforms for investors can provide firm-specific information to the best of our knowledge. Our work builds on these differences and provides further evidence that investors' disagreement on StockTwits provides firm-specific information that can predict financial markets. Our study is closely related to that of Ding et al. (2019), who used the number of articles published on SeekingAlpha to predict return synchronicity. Our study provides evidence that, besides attention, social interactions among investors on social media platforms can predict return synchronicity. In relation to these findings, we are the first to provide direct evidence that disagreement on StockTwits provides firm-specific information that can be useful for a broad range of stakeholders, mainly retail investors and portfolio managers in financial markets.

The rest of the paper is organized as follows: Section 2 discusses the literature review and hypothesis development, Section 3 describes the data and explains the research design, Section 4 discusses the empirical results, Section 5 presents robustness checks, and Section 6 presents the conclusion.

3.2. Literature Review and Hypothesis Development

3.2.1. Background on Return Synchronicity

Return synchronicity is defined as the extent to which stock return comoves with the market and industry returns (Durnev et al., 2003). Therefore, higher return synchronicity means stock returns are explained by industry and market returns, and low synchronicity means the variation in stock returns has a weak association between market and industry returns. Roll (1988) provides corroborative evidence of the weak association between stock returns and industry and market movements. He further suggests that this weak association is attributed to firm-specific information incorporated in the stock prices. Similarly, Shiller (1989) presents further evidence that UK and US firms' dividends cannot fully explain the comovement of stock prices between two countries. To test whether less return synchronicity provides firm-specific information, Durnev et al. (2003) provide evidence that firm-specific stock price variability is positively correlated with price informativeness measures. Thus, less synchronicity provides higher firm-specific information.

Building on the exciting research agenda of Roll (1988), further studies have confirmed that various factors determine return synchronicity. In this vein, Pindyck and Rotemberg (1993) highlight the role of market segmentation partly explained by firm size and institutional ownership to describe the individual stock comovement; Morck et al. (2000) present a cross-country sample showing the negative association between return synchronicity and government protection of property rights[43]; Durnev et al. (2004) argue that less synchronicity (higher firm-specific information) is associated with the efficient allocation of capital by the firms; Chan and Hameed (2006) provide evidence that higher analyst coverage is associated with higher return synchronicity, suggesting a lower inflow of firm-specific information by the analysts; Jin and Myers (2006) argue that a lack of transparency increases return synchronicity; and An and Zhang (2013) present a negative association between return synchronicity and crash risk, among others.

Although previous studies provide substantial evidence that less return synchronicity provides firm-specific information, Barberis et al. (2005) provide evidence against this fundamentals-based view. Based on univariate analysis, they argue that additions in the S&P 500 index suggest that a higher level of firm-specific stock variation is synchronized with market movements and has nothing to do with firm-specific information. In similar lines, Green and Hwang (2009), using stock splits, present evidence supporting Barberis et al. (2005), suggesting that the firm-specific information cannot explain price-based comovement. Conversely, Chen et al. (2016) provide corroborative evidence that the inclusion of momentum stocks likely explains some of the sample results reported by Barberis et al. (2005) and Green and Hwang (2009). This is because, after univariate analysis, factor loading based on Dimson (1979), and matching techniques, beta changes are indistinguishable before and after index additions and stock splits. These studies highlight the need to further investigate the relationship between return synchronicity and price informativeness using better identification strategy and datasets.

3.2.2. Return Synchronicity and Firm Information Environment

The firm information environment plays a pivotal role in determining the flow of firm-specific information in the financial markets. In this vein, Piotroski and Roulstone (2004) argue that, contrary to the existing notion that higher analyst coverage provides firm-specific information, higher analyst coverage exacerbates market information flow, and Jin and Myers (2006) provide evidence that less transparency is associated with higher return synchronicity.

[43] This negative association of property rights is also consistent with Roll (1988) since weak enforcement of such rights is likely to impede firm-specific informed trading.

Veldkamp (2006) argues that when investors consume information from a common source, information about one asset affects the price of other assets. Therefore, this search for information makes asset prices more efficient and causes some assets to comove with other assets. One of the key challenges highlighted by Veldkamp (2006) when seeking to understand information-driven return synchronicity is the lack of data on proxy investors' information.

Recent studies by Ding et al. (2019) using the number of articles on the social media platform argue that social interaction among investors increases the flow of firm-specific information in financial markets. Drake et al. (2017) offer compelling evidence that the comovement of investors' attention has significant consequences on the comovement of returns between the firm and its peers. Jiang et al. (2019) present evidence that co-discussed stocks are actively traded and have higher comovement with the stocks which are discussed together. Similarly, Shiyang Huang et al. (2019) provide unique evidence that investors pay more attention to market-level information due to scarce cognitive resources. In the presence of attention shocks, market-level information increases the marginal utility of information consumed by investors.

Although these studies provide compelling evidence that attention plays a pivotal role in determining stock comovement, Shiller (1992) suggests that investing is a social activity in which investors share their opinions about different investment approaches[44]. Moreover, recent studies by Hirshleifer and Teoh (2009) and Hirshleifer (2015) suggest that behavioral and psychological aspects of investors' decision-making are based on current information and priors. Therefore, these studies warrant further evidence to investigate the association between social interactions and firm-specific information in financial markets.

3.2.3. The Role of Disagreement in Financial Markets

Disagreement arises due to differences in opinion. However, it is pertinent to ask whether disagreement converges or if investors agree to disagree with each other based on the assumption that all investors have heterogeneous priors and the ability to interpret new information in an entirely different fashion (Harris & Raviv, 1993). In this research line, Aumann (1976) argues that convergence in disagreement occurs when investors have common priors and a shared understanding of each other's posterior beliefs. However, later studies

[44] Hong et al. (2004) provide evidence consistent with Shiller (1992) and argue that social interactions among investors increases the stock market participation.

provide contrasting evidence that disagreement persists in financial markets due to the quality of information signals and investors' uncertainty to interpret them (Acemoglu et al., 2006; Kandel & Pearson, 1995; Varian, 1985).

It is also important to note that the convergence of disagreement depends entirely on the probability of learning from investors' information signals. However, there is ample evidence of investors' inconsistent learning in financial markets (Banerjee et al. (2019), e.g., investors' over- and underreactions (Barberis et al., 1998) to prices, investors' overconfidence (Eyster et al., 2019; Odean, 1998), excess volatility, and returns in financial markets. Overall, there is a consensus in the literature that differences in opinion drive financial markets. Disagreement does not converge as investors have heterogeneous priors, and the instantaneous flow of information in financial markets induces investors to continuously update their beliefs (Banerjee & Kremer, 2010).

One of the critical challenges in financial markets is to find a suitable proxy for investors' disagreement. In this regard, previous studies have commonly used analyst dispersion (Diether et al., 2002), returns (Cen et al., 2017), and options (Golez & Goyenko, 2019). Similarly, with the emergence of social media platforms for investors, recent studies have also used investors' disagreement on StockTwits (Al-Nasseri & Menla, 2018; Giannini et al., 2019). A recent study by Cookson and Niessner (2019), using data from StockTwits and measuring investors' disagreement within and across groups (investment philosophies), argues that investors' disagreement arises due to differences in investment approaches and investors' experience. These findings are consistent with Kandel and Pearson (1995) and offer insightful evidence by providing a unique proxy for investors' disagreement using social media platforms for investors.

Based on the existing literature, there is a consensus that investors' disagreement plays a vital role in financial markets.[45] For example, Fama and French (2007) revised their assumption about the distribution of future payoffs and suggested that investors' disagreement affects asset prices. However, current literature does not address the extent to which investors' disagreement on social media platforms for investors provides firm-specific information on financial markets. In this line of research, most studies have mainly focused on predicting volumes and

[45] Sheng Huang et al. (2019) concluded that management–investor disagreement can influence management to replace CEOs. In another study, Ayotte (2020) argue that firms have an incentive to exploit disagreement among investors by reducing borrowers' cost of funds and directly affecting the capital structure of the firm.

returns based on investors' sentiments on Internet discussion boards (Antweiler & Frank, 2006) such as Twitter (Bollen et al., 2011; Sprenger et al., 2014), SeekingAlpha (Campbell et al., 2019; Chen et al., 2014), and StockTwits (Renault, 2017). This study fills this research gap by investigating the role of disagreement on investors' social media platforms and predicting return synchronicity.

3.3. Hypothesis Development

Standard asset-pricing models assume that financial markets can process new information at a sufficient speed to keep pace with information arrivals in financial markets. Consequently, due to the instantaneous flows of information, asset prices adjust according to the available information. This assumption is consistent with rational expectation models in frictionless markets. However, in a market with friction and limits on arbitrage, information friction, endowed biases among investors and various stakeholders in financial markets, and scarce cognitive resources challenge investors' ability to process public and private signals.[46] A substantial body of literature focuses on understanding how investors process information in financial markets. For example, Daniel et al. (1998) discuss the role of psychological factors to explain the under- and overreactions of investors to specific information signals, belief heterogeneity, and differences of opinion (Banerjee et al. (2009).

The disagreement hypothesis is based on three main assumptions explained by Hong and Stein (2007) disagreement model. The first is the flows of information in financial markets, suggesting that even in the presence of a continuous stream of information, not all investors consume information at the same level due to scarce cognitive resources. The second is that investors have limited attention due to scarce cognitive resources. This assumption suggests that information released in an attention-grabbing[47] manner will have more implications for investors than general news flowing into the market. The third assumption suggests that investors have heterogeneous priors, and while all investors may receive the same information simultaneously, they interpret that information based on their heterogeneous priors. The disagreement model offers a natural framework for understanding disagreement in financial markets, explaining mechanisms that can generate investor disagreement. In line with Hong

[46] Eyster et al. (2019) propose a cursed trader's model based on private information acquired by other investors in financial markets. They conclude that cursed traders actively trade and assume excessive risk in financial markets. propose a model of costly information, whereby investors' interpretation of prices is noisy due to bounded rationality.
[47] Attention-grabbing features of information signals include categories such as the salience of information signals.

and Stein's (2007) disagreement model, we argue that, due to the instantaneous flows of information in financial markets via different information channels, investors on social media platforms such as StockTwits consume information from every available information channel and participate in discussions on StockTwits using cashtags. Those investors interpret information signals based on their heterogeneous priors[48] and participate in discussions by sharing their ideas and recommendations. Since investors will only update their recommendations based on their interpretation of information signals on StockTwits, frequent updates to investors' recommendations will increase disagreement between investors. Increased disagreement among those investors offers a unique opportunity to understand the information acquisition process in financial markets and the social benefits resulting from the interactions between investors on such platforms. In line with these arguments, we expect that disagreement on StockTwits adds additional value to investors' economic models by offering firm-specific information. This leads to our main hypothesis:

H1: *Ceteris paribus, greater disagreement among investors on StockTwits results in a decrease in return synchronicity.*

3.4. Data Description and Research Design

3.4.1. Data Description

Our firm-level data come from NYSE/AMEX/NASDAQ-listed companies' common stocks with share codes of 10 and 11 from January 2013 to December 2017. Stock data are collected from the Centre for Research in Stock Prices (*CRSP*). Quarterly firm-level financial statement data is collected from Compustat, and I/B/E/S and insider-trading data are collected from Thomson Reuters. We download firm-level media coverage data from Ravenpack News Analytics (RPNA) services. RPNA offers greater flexibility and in-depth media coverage as compared to any other news database. Overall, we download 2.2 million news articles with daily coverage of our sample firms. The daily news articles are aggregated monthly to calculate media coverage. To ensure that we only download relevant news data, we apply RPNA *Relevance* and *Novelty* filters.[49] The average *Relevance* score for our media coverage data is

[48] Harris and Raviv (1993) and Kandel and Pearson (1995) call these heterogenous priors economic models.
[49] Ravenpack news analytics standardizes Relevance and Novelty scores from 0–100. In the case of relevance, a score of 0–100 indicates how strongly a firm is related to a news story; values greater than 50 suggest that the firm is discussed along with other firms, and values greater than 75 suggest the firm is discussed significantly in the news. In the case of novelty, a score of 0–100 indicates how novel the news story is within a given 24-hour period. Values greater than 50 suggest that multiple stories exist within the 24-hour period, and values equal to 100 suggest that the news story is completely novel within the same period.

89.90, with a median of 98 and a standard deviation of 15.45. Similarly, the average *Novelty* score is 94.03, with a median of 100 and a standard deviation of 12.42. Overall, the average number of news articles is 2.95, with a median of 2.94 and a standard deviation of 1.06 articles per month. To further ensure data relevance and exclude small and less traded firms, we only include firms with an average share price above $1, and whose trading data for the last one year are available in CRSP. We then match our sample firms with the StockTwits sample. Based on data availability while matching with different databases, we have 956 firms in our final sample with 53,778 firm-month observations from 2013 to 2017. The overall average *Firm Size* is $19.33 billion with a median of $3.21 billion, and the average *Analyst Coverage* is 12 analysts for each firm.

To construct disagreement as our main variable of interest, we collect data from StockTwits, a popular social media platform for investors. StockTwits is by far the largest social media community for investors and traders, with more than 3 million members, 5 million monthly messages, and 3 million monthly visitors.[50] The main user interface of StockTwits is user-friendly; investors can post Twitter-like messages up to a 140-character limit.[51] One of the distinguishing features of StockTwits is investors' user profiles, where any investor can volunteer to disclose their asset choices, investment approaches, and investment term preferences. Moreover, investors can create a customized watchlist to view StockTwits' ideas directly relevant to their investment preferences. To collect data from StockTwits, we developed a python program to connect with StockTwits API and collect multiple data points based on numerous iterations from January 2012 to December 2018. StockTwits relies on cashtags (for example, $AAPL) as company identifiers. We were able to harvest more than 38 million ideas during our data collection, posted by approximately 297,000 users, discussing 8,394 companies listed in the US and other national stock exchanges.

We apply several filters to ensure the quality of data harvested from StockTwits. To this end, to ensure that data are relevant, we only keep data from January 2013 to December 2017. To avoid concerns related to shared attention and discussion of multiple cashtags in a single message, we only keep StockTwits ideas that contain single cashtags and only discuss one company in a given message. Our final filter restricts our StockTwits sample to those companies listed in NYSE/AMEX/NASDAQ as common stocks. This filter is essential in

[50] https://about.stocktwits.com/

[51] On May 8, 2019, StockTwits increased their character limit to 1000 characters. However, this occurred outside our sample period.

terms of investors' ability to identify companies and discuss their ideas uniquely. After applying these filters, we are left with approximately 16 million ideas containing 1,890 unique cashtags. We match the StockTwits sample with our US firm-level data, resulting in 956 firms and 53,778 firm-month observations in our final sample, with more than 12 million ideas.

[Insert Table 3.1 here]

3.4.2. StockTwits: Where Is It All Coming From?

StockTwits is a social media platform for investors. With the recent surge in information flows in financial markets, it is crucial to understand that ideas expressed on StockTwits must be accurate representations of investors' opinions. We use StockTwits for the following reasons. First, the heterogeneity of investors is a distinct attribute of StockTwits. Investors can voluntarily disclose their experience level (professional, intermediate, or novice) and can disclose their investment approaches. Second, StockTwits is free for anyone wishing to share and post their ideas. This motivates investors to engage in discussions with other investors. Third, during this information-sharing process, investors have the opportunity to increase their social influence by increasing their number of followers after sharing unique investment ideas and expert analysis. Therefore, StockTwits encourages investors to share their ideas and learn from investment professionals who are experts in their fields.

It is also vital to understand the dynamic structure of StockTwits compared to other social media platforms for investors[52]. In a recent study, Cookson and Niessner (2019) argue that investors on StockTwits are less inclined to post fake news and more interested in posting reliable information to become famous. Clarke et al. (2020), using data from SeekingAlpha, differentiate between fake news and legitimate news articles. They argue that although fake news articles attract investors' attention, legitimate news articles generate higher trading volumes. However, it is pertinent to note that investors' motivation to post on StockTwits as compared to SeekingAlpha is entirely different. On SeekingAlpha,[53] authors are paid for their articles if they start to attract a certain number of readers; on StockTwits, users post without

[52] Unlike other social media platforms, StockTwits is an investor-oriented social media platform. It offers investment/trade specific recommendations' options as well as catogorizing users into various type of investors.
[53] SeekingAlpha payment terms and conditions can be found here: https://seekingalpha.com/page/payment-terms

seeking any monetary benefit.[54] Therefore, unlike any other social media platform, ideas posted on StockTwits are less subject to potential bias.

Next, misinformation does not seem to be a matter of concern in our data set for the following reasons. First, it is highly unlikely that investors on StockTwits can influence financial markets by sharing false information. This is because several other information channels will have a crowding-out effect on these investors' incorrect information. Second, investors' primary motivation to share opinions on such platforms is to gain popularity on them, and for that reason, disseminating false information and fake news can be harmful to their StockTwits profile. Third, there are no financial benefits for investors who post their ideas on StockTwits, except sharing ideas and recommending stocks. More importantly, unlike Twitter, StockTwits is not seen as a marketing platform for individuals. Finally, since our sample firms have large market capitalization and high liquidity, it is less likely that misinformation (if any) on StockTwits can influence these stocks' prices.

Our StockTwits sample contains more than 12 million ideas,[55] posted by 162,836 distinct investors[56]. Panel A of Table 3.1 presents summary statistics for the StockTwits data on our sample firms. All StockTwits variables are log-transformed, except disagreement. Overall, the average disagreement on StockTwits is 0.53. The average monthly frequency of ideas on StockTwits is 3.99, posted by 3.75 investors on average, with an average social media experience[57] of more than 13 months. To better understand the impact of an extensive social network, we constructed *Network* as a monthly variable by aggregating the number of followers of distinct investors who post ideas on StockTwits related to specific firms. The average monthly *Network* level is 12.97, with a median of 13.28. We also calculate the number of revisions a distinct investor makes on the same stock. *Revisions* are defined as the sum of the number of times per day a distinct investor revises their recommendations (e.g., Bullish to Bearish or vice versa), which is then aggregated at a monthly frequency. *Revisions* are log-transformed; the average number of *Revisions* is 3.61, and the median is 3.22.

---

[54] On April 10, 2017, the SEC cracked down on alleged stock-promotion schemes on SeekingAlpha, whereby some authors were paid to write in favor of certain companies listed in the United States. Further details can be found here: https://seekingalpha.com/article/4061813-seeking-alpha-applauds-secs-actions-to-stomp-out-stock-promotion
[55] StockTwits messages posted by investors are referred to as ideas on StockTwits.
[56] We recognize all users who post ideas on StockTwits as investors.
[57] The average social media experience is calculated as the number of months between the date of joining StockTwits until the investors posted their first idea in our sample.

To share ideas on StockTwits, investors must create a user profile. Fig. 3.1 presents a summary of information from investors' profiles on StockTwits. Panel A of Fig. 3.1 presents the distribution of investors according to their StockTwits joining year. The x-axis is the investors' joining year, and the y-axis is the percentage of investors joining StockTwits in that specific year. It is pertinent to mention that StockTwits has become more popular in recent years. Panel B shows the distribution of ideas and investors in the sample years, where the x-axis is the sample year. The y-axis is the percentage of investors/ideas on StockTwits discussing the sample firms. It is important to understand the distribution of StockTwits ideas and investors at the sector level since some industries on social media receive more coverage than others. In this regard, Panel C presents the distribution of ideas and investors across different sectors based on the Global Industrial Classification System (GICS) four-digit sectors. It is evident that the healthcare sector has the highest concentration of investors, and consumer discretionary and information technology have the highest concentration of ideas, respectively.

[Insert Figure 3.*1 Here]*

Due to the breadth of information in our data harvested from StockTwits, we try to understand the geographical distribution of ideas and investors in the US at the state level.[58] To do this, we standardize users' locations at the city and state levels so that we can allocate ideas and investors to relevant US states using state-level coordinates obtained from the 2017 US Census website geographic data. Fig. 3.2 presents the distribution of ideas and investors across US states. Their geographical distribution is important for the following reasons. First, it shows that most of the investors who frequently post on StockTwits are based in the USA, with only 2% of investors and ideas coming from the rest of the world.[59] Second, the geographical distribution of investors suggests that the majority of investors come from the three largest states, i.e., California (16.18%), New York (16.17%), and Texas (8.32%). A list of randomly selected ideas is presented in Appendix 3.3.

[Insert Figure 3.*2 Here]*

[58] Sixty-two percent of investors out of 162,836 disclosed their location in their StockTwits public profiles. We also lose some investor-level location data in the data-cleaning and standardization process. For example, we cannot map investors who only disclosed their country as their location.
[59] As compared to country-level information disclosed by investors on their StockTwits public profiles.

3.4.3. Variables Construction

3.4.3.1. Return Synchronicity

The measure of return synchronicity is calculated based on the value of the coefficient of determination (R^2). To derive the value of R^2, we use Carhart (1997) four-factor model, since the momentum factor, in addition to the Fama–French factors, which may also be a source of variation in return synchronicity, can better account for idiosyncratic risk. We use the following equation to calculate the value of R^2:

$$Ret_{i.d} = \beta_0 + \beta_{mkt.i}\, MKT_d + \beta_{HML.i}\, HML_d + \beta_{SMB.i}\, SMB_d + \beta_{UMD.i}\, UMD_d + \varepsilon_{i,d} \tag{1}$$

Where $Ret_{i.d}$ is the daily return on stock *i* at day *d*. The right side of the equation is market (MKT), high minus low (HML), small minus big (SMB), and momentum (UMD) factors[60]. To ensure the availability of sufficient numbers of daily observations to calculate monthly R^2, the stock return must have at least 50% of non-missing observations on trading days in a given month. $R^2_{i,t}$ derived from Eq. (1) is the coefficient of determination, ranging between 0 and 1 for stock *i* during month *t*. However, the existing measure has a high level of skewness and kurtosis, resulting in some econometric issues. To deal with this problem, we take the natural logarithm of $R^2_{i,t}$, which is consistent with the existing literature (Morck et al., 2000; Piotroski & Roulstone, 2004). Our final measure presents an unbiased proxy for monthly return synchronicity $Sync_{i,t}$ for stock *i* during month *t* and is calculated as follows:

$$Sync_{i,t} = \ln\left(\frac{R^2_{i,t}}{1 - R^2_{i,t}}\right) \tag{2}$$

Summary statistics for $Sync_{i,t}$ are presented in Panel C of Table 3.1. The average *Return Synchronicity* is –0.37, with a standard deviation of 1.04 and a median of –0.33. To further check the robustness of our synchronicity measure, we also calculate the return synchronicity following Roll (1988), who argues that market and industry returns, as well as a firm's stock returns, are inversely related to the firm-specific information incorporated in stock prices.

[60] The factors' return data is downloaded from Kenneth R. French Website.

Following Peng and Xiong (2006) and Antón and Polk (2014), we also use Pearson's correlation coefficient to measure return synchronicity, which is the correlation between firm return and market return. We also use alternative methods to calculate return synchronicity and our results remain consistent.

3.4.3.2. Recommendation Classification Model

Investors' recommendations on StockTwits represent voluntary disclosures. Therefore, not all ideas posted on StockTwits contain investors' recommendations. To construct our measure of disagreement, we use supervised machine-learning classification models to predict investors' recommendations. To this end, a key requirement is to build a robust training data set that is sufficiently large and accurate. Unlike previous studies that mainly rely on hand-classified training data of up to 3000 posts, we use more than 1 million pre-classified ideas labeled by investors on StockTwits. Therefore, our training data set is both large and accurate. We use Baziotis et al.'s (2017) Ekphrasis library for data pre-processing, which is a specialized text pre-processing[61] tool for online social networking platforms. There is no rule of thumb for choosing classification models. It mostly depends on the type of data set and prediction accuracy after cross-validation tests. We use the Random Forest Decision Trees (RFDT) method to more quickly implement and more easily interpret, based on decision trees (compared to the Support Vector Machine). For example, Fernández-Delgado et al. (2014) tested 179 classifiers and concluded that the Random Forest model is one of the best classification models with a close match for Support Vector Machine (SVM) models. Therefore, in our study, we use the RFDT model for recommendation classification. However, to further ensure our results' comparability and robustness, we also use the SVM and Maximum Entropy (MaxEnt) models.[62]

The recommendation classification process is completed in two steps. In the first step, we create multi-way decision trees from the data set such that the data set is split into smaller subsets to predict target values. To maximize information gain and reduce the level of uncertainty in predictions, we use entropy as our impurity criterion. During the prediction process, conditions are presented as nodes, and possible outcomes are presented as edges. The decision-trees method is advantageous because of its quick application and faster turnover on large training data sets. However, one of the main drawbacks of using decision trees is

[61] Further details on data pre-processing are presented in Appendix 3.1.
[62] Further discussion is presented in Section 6, along with regression results based on Eq. (5).

overfitting. Although tree depth is vital to address the overfitting problem, we use the Random Forest[63] (RF) model in the second step. The Random Forest model operates as an ensemble and uses these decision trees to pool all classifications and predict final recommendations. Our final classification results are based on feature selection and are robust under tenfold cross-validation, with an F1 score of 89% and an overall accuracy of 82%.

3.4.3.3. Disagreement

To construct our disagreement measure, we follow Antweiler and Frank (2004) approach to calculate the average recommendations as follows:

$$Rec_{i,t} = \frac{Rec_{i,t}^{Bullish} - Rec_{i,t}^{Bearish}}{Rec_{i,t}^{Bullish} + Rec_{i,t}^{Bearish}} \in [-1, 1] \tag{3}$$

Where $Rec_{i,t}$ is the average recommendation for firm *i* in month *t*. Similarly, $Rec_{i,t}^{Bullish}$ is the aggregate bullish recommendations, and $Rec_{i,t}^{Bearish}$ is the aggregate bearish recommendations. Our average recommendations range between –1 and 1. Our disagreement measure deviates from Antweiler and Frank (2004) assumption, as they assume latent disagreement when there are no posts. In this regard, our approach is consistent with Cookson and Niessner (2019), who assume that no posting means no disagreement. For this purpose, we normalize no-posting cases equal to 0. Therefore, we calculate the overall disagreement as follows:

$$Disagreement_{i,t} = \sqrt{1 - Rec_{i,t}^{2}} \in [0, 1] \tag{4}$$

$Disagreement_{i,t}$ in Eq. (4) is the overall disagreement among investors' recommendations for firm *i* in month *t*. The value of $Disagreement_{i,t}$ ranges between 0 and 1, where 0 represents the complete agreement and 1 represents the complete disagreement between investors on StockTwits. Panel A of Table 3.1 presents overall disagreement with an average of 0.53, a standard deviation of 0.56, and a median of 0.59.

[63] Model derivation is presented in Appendix 3.1.

3.4.4. Research Design

To examine the relationship between Return Synchronicity and *Disagreement*, we estimate the following model:

$$
\begin{aligned}
Sync_{i,t} = {} & \beta_0 + \beta_1 Disagreement_{i,t} + \beta_2 Media\ Coverage_{i,t} + \beta_3 Analyst\ Coverage_{i,t} \\
& + \beta_4 Leverage_{i,t} + \beta_5 Adv/Sales_{i,t} + \beta_6 \,{}^{Market}/_{Book}\, Ratio \\
& + \beta_7 Firm\ Size_{i,t} + \beta_8 ROA_{i,t} + \beta_9\, Earnings\ Volatility_{i,t} \\
& + \beta_{10} Sales\ Growth_{i,t} + \beta_{11} Real\ GDP_{m-1} + V_i + V_t + V_p + \varepsilon_{i,t}
\end{aligned}
\tag{5}
$$

Where $Sync_{i,t}$ is the *Return Synchronicity* of firm i at time *t*. The key explanatory variable in Eq. (5) is the $Disagreement_{i,t}$ among investors on StockTwits about firm i at time *t*. We estimate the equation using the fixed-effects estimator to account for unobserved firm-specific heterogeneity (V_i). We also control for time (V_t) and industry (V_p) fixed effects by including month and industry dummies, capturing time-varying and industry-specific movements. Following Petersen (2009), we use two-dimensional clustering at firm and time levels to account for any within-group correlation that may influence standard errors.[64]

We also employ firm-level and market-level control variables, which may directly or indirectly affect our variables of interest. For example, we use *Media Coverage* to control the effect of firm-specific information from alternative sources of information. Following Chan and Hameed (2006), to control for the effect of analysts following the firm, we use *Analyst Coverage*. Firm Size is used to control the firm's size to attract investors' attention, and because the demand for analysts' recommendations is directly proportional to the firm's size. To account for the creditors' monitoring role, which may influence the firm-information environment, we use *Leverage*. Firth et al. (2008) argue that firms with higher leverage have higher monitoring by creditors, consequently playing a vital role in managers' decision-making. We use firms' *Adv/Sales*[65] and *Market/Book* ratios to control the effect of firm influence on the external information environment and control the effect of firm-level growth opportunities based on the market value of equity, respectively. Similarly, we use *ROA*, *Earnings Volatility,* and *Sales*

[64] Our results remain consistent when clustering at the firm level only.

[65] Chemmanur and Yan (2019) present that an increase in advertising increases the firm's visibility among investors and attracts investors' attention. In another study, Grullon et al. (2004) present that firms with higher advertising costs have higher liquidity and a large investor base.

Growth as indirect proxies for investors' attention in financial markets. Finally, to control for the effect of macroeconomic trends, we use the monthly lagged value of *Real GDP* (Brockman et al., 2010). To make the statistics intuitive, we standardize all the right-hand-side variables in our regression. Variable definitions are presented in Appendix 3.2.

3.5. Empirical Analysis

3.5.1. Correlation Table

To understand the relationship between explanatory variables and *Return Synchronicity*, we measure Pearson's correlation coefficient between all variables.[66] The correlation between *Return Synchronicity* and *Disagreement* is statistically significant and negative, consistent with our hypothesis, according to which disagreement might reduce return synchronicity. *Analyst coverage* is an important source of industry and market-level information, and it has a 23.2 and 17.7% correlation with *Return Synchronicity* and *Disagreement,* respectively. *Media Coverage* is another variable of interest since it plays a pivotal role in influencing firms' information environment. It has a 7.4% correlation with *Return Synchronicity* and a 17.5% correlation with *Disagreement*. Overall, the correlation between *Return Synchronicity* and the remaining explanatory variables is less than 15%, suggesting no multicollinearity problem in our regressions.

3.5.2. Main Results

The results presented in Table 3.2 are estimated using Eq. (5). We use ordinary least square regression (OLS) with the firm, month, and industry fixed-effects as our main regression model. To understand the impact of *Disagreement* on *Return Synchronicity,* Model (1) shows the impact of *Disagreement* without controlling for the effects of firm-level covariates and macroeconomic trends. The result supports our hypothesis that *Disagreement* reduces *Return Synchronicity.* Considering that we standardized the right-hand-side variables, a standard deviation increase of one in *Disagreement* results in 6.9% greater inflows of firm-specific information into financial markets, which is economically significant. One of the key factors that may influence firm-specific information flows in financial markets is *Media Coverage.*[67] Therefore, in Model (2), we add *Media Coverage* to account for the impact of firm-specific

[66] For brevity reasons, the table is presented in the online Appendix.
[67] To account for the aggregate effect of media coverage. Later in our analysis, we present the impact of media coverage after dissecting it based on news types, news topics, and news sources.

news in our model. Our results show that there is a negative relationship between *Media Coverage* and *Return Synchronicity,* suggesting that higher media coverage results in greater inflows of firm-specific information into financial markets.[68] Similarly, the *Disagreement* coefficient remains significant and negative, suggesting that investors actively use social media platforms such as StockTwits and consume information from such platforms in addition to conventional media coverage. The magnitude of the impact of *Disagreement* on *Return Synchronicity* in Model (2) suggests that a standard deviation increase of one in *Disagreement* results in 5.4% higher inflows of firm-specific information in financial markets.

[Insert Table 3.2 here]

In Model (3), we add firm-level covariates and macroeconomic trends to control for the effects of firm-level factors and macroeconomic trends influencing firm-specific information in financial markets. Specifically, we add *Leverage* and *Analyst Coverage* to control for the effect of the firm-information environment. The results from *Leverage* are consistent with Armstrong et al. (2010), suggesting that higher *Leverage* will increase the monitoring benefits for shareholders. The results for *Analyst Coverage* suggest that greater analyst coverage means more focus from analysts on the mapping of the industry- and market-level information instead of firm-specific information. These results are consistent with Piotroski and Roulstone (2004) and Chan and Hameed (2006), who concluded that *Return Synchronicity* is positively associated with *Analyst Coverage*. We use the firms' *Adv/Sales* ratio to control the firm's flows of firm-specific information to increase the firm's visibility and attract individual and institutional investors. We use *Firm Size* to control the size of firms to influence the firm-information environment. Da et al. (2011) suggest that several indirect proxies of attention may affect information flows in financial markets. To account for such factors, we use firm-level earnings (*ROA*), *Earnings Volatility,* and *Sales Growth* as indirect proxies for investor attention, which may influence the firm-information environment. Following Chue et al. (2019) and Brockman et al. (2010), we add the lagged value of $Real\_GDP_{t-1}$ to Model (3) to account for macroeconomic trends and their implications for *Return Synchronicity*.

Our results in Model (3) remain consistent after controlling for the effects of firm-level covariates and macroeconomic trends, suggesting that one standard deviation increase in *Disagreement* results in 5.7% higher inflows of firm-specific information in financial markets.

[68] These findings are consistent with the recent study by Dang et al. (2020), who suggest that firm-level media coverage provides firm-specific information to investors in financial markets.

To further test the validity of our results, we run OLS regression in Model (4) without using any fixed effects. Our results remain consistent in Model (4). Similarly, to account for any cross-sectional correlation and estimate consistent standard errors, we use Fama and MacBeth (1973) two-step regression. The results presented in Model (5) remain consistent with our previous findings.

3.5.3. Evidence of Firm-Specific information from Disagreement

In this section, we explicitly test our assumption that disagreement on StockTwits provides firm-specific information to investors in financial markets. For this purpose, we use different identification strategies (*Price Informativeness*, *Recommendation Revisions,* and *Media Coverage*) to explore the role of disagreement in predicting firm-specific information.

3.5.3.1.Price Informativeness

According to Roll (1988), a lower value of R^2 is due to arbitrageurs who gather and possess private information while trading in financial markets. This is because he could not find any association between firm-specific[69] stock-price movements and news releases, and he further suggests that "The financial press misses a great deal of relevant information generated privately" [p. 564]. In our setting, we argue that investors' social media platforms offer a unique opportunity for investors to share their ideas. Under this assumption, if investors on StockTwits are discussing firm-specific information, the incorporation of price informativeness should illustrate the impact of disagreement to predict future earnings. We follow Ayers and Freeman (2003) model to test whether *Disagreement* on StockTwits improves the prediction accuracy of future earnings of firms and estimate the following regression equation:

$$
\begin{aligned}
CAR_{i,t} = \beta_0 + \beta_1 Disagreement_t + \beta_2 \Delta Earnings_{t+1} + \beta_3 Disagreement_t * \Delta Earnings_{t+1} \\
+ \beta_4 \Delta Earnings_t + \beta_5 Disagreement_t * \Delta Earnings_t + \beta_6 \Delta Earnings_{t-1} \\
+ \beta_7 Disagreement_t * \Delta Earnings_{t-1} + \beta_8 Controls + \varepsilon_{i,t}
\end{aligned}
$$

[69] Durnev et al. (2003) test this hypothesis and provide further evidence that greater firm-specific price risk is associated with higher stock-price informativeness. This is consistent with Roll (1988) argument that firm-specific information reflects the presence of arbitrageurs capitalizing based on private information in financial markets.

All variables are calculated at the quarterly frequency to maintain consistency between *Disagreement* and earnings variables.[70] $\Delta Earnings$ is calculated as the change in earnings of firm *i* from quarter *t-1* to *t* scaled by the market value of the equity of firm *i* at the beginning of the quarter.[71] $\Delta Earnings_{t+1}$, $\Delta Earnings_{t}$, and $\Delta Earnings_{t-1}$ are proxies for lead, contemporaneous, and lag changes in earnings, respectively. (6)

$CAR_{i,t}$ is the cumulative abnormal return of firm *i* calculated as the sum of daily abnormal returns for quarter *t.* Abnormal return is calculated as follows:

$$AR_{i,t} = R_{i,t} - R_{SBM,t}$$

Where $AR_{i,t}$ is the abnormal return for firm *i* in each quarter *t*, $R_{i,t}$ is the daily return for firm *i* and $R_{SBM,t}$ is the equal-weighted 5 × 5 portfolio return calculated using size and Book/ Market value of equity *(BE/ME)* of firm *i*.

Our main variables of interest are the interactions between *Disagreement* and $\Delta Earnings_{t+1}$ and between *Disagreement* and $\Delta Earnings_{t-1.}$ The intuition is that if low return synchronicity implies more firm-specific information is reflected in prices, we would expect that increased *Disagreement* can increase price leads of earnings. If disagreement among investors reflects noise, disagreement should go in the opposite direction for post-earnings announcement drift (*PEAD*). We use the impact of $\Delta Earnings_{t+1}$ and $\Delta Earnings_{t-1}$ on abnormal returns to identify price leads of earnings and the post-earnings announcement drift.

[Insert Table 3.3 *here]*

Table 3.3 presents the regression results estimated using Eq. (6). We use OLS regression with firm-level fixed effects as our main model and then test the robustness of our results using simple OLS and Fama–MacBeth regressions, respectively. Model (1) presents the stand-alone regression results where only contemporaneous earnings are positive and statistically significant at the 1% level, suggesting that security price responds to contemporaneous earning rather than lead and lag changes in earnings. Model (2) presents the regression results for the interaction between *Disagreement* and change in earnings at the leading, contemporaneous,

[70] Ayers and Freeman (2003) use annual data instead of quarterly data. However, social media platforms for investors post ideas based on a short-term approach. In this regard, Giannini et al. (2019) present evidence that investors’ opinions either converge or diverge around earnings announcements, without having any implications in the long run.
[71] We use Compustat data item 18 (income before extraordinary items) as a measure of earnings.

and lagged levels. If the increase in price informativeness illustrates the impact of *Disagreement* on incorporate future earnings, then β_3 should be positive. Our regression results in Model (2) present that coefficient β_3 is positive and statistically significant at the 1% level, which is consistent with our projection that security prices reflect future earnings when there is more *Disagreement* among investors. In contrast, the coefficient β_7 is negative and statistically significant at the 1% level, suggesting that disagreement among investors does not reflect noise. *Disagreement,* in fact, reduces the price response to $\Delta Earnings_{t-1}$. These results are in line with our hypothesis that *Disagreement* increases price informativeness; firms with high levels of *Disagreement* have lower post-earnings announcement drift (*PEAD*).

Next, the interaction between $\Delta Earnings_t$ and *Disagreement* is negative but insignificant, suggesting that *Disagreement* has no impact on price responses to contemporaneous earnings. It is pertinent to note that the insignificant interaction between contemporaneous earnings and *Disagreement* implies that the lead effect of *Disagreement* is not due to changes in the magnitude of price responses to earnings. If firms with more *Disagreement* have higher earnings persistence than others, we should also find a positive association between contemporaneous earnings and *Disagreement.* In our case, the coefficient from β_5 is insignificant, thus further strengthening our results from β_3. Furthermore, we include firm size as a control in Model (3). Our results remain consistent when we add the control variable.

Models (4) to (6) present the regression results without using firm-level fixed effects. Finally, in Models (7) to (9), we present our results using Fama–MacBeth regression, as in Ayers and Freeman (2003). Overall, our results remain consistent, and the coefficient β_3 is positive. These findings suggest that price leads increase as *Disagreement* increases. Therefore, *Disagreement* on StockTwits is an important source of firm-specific information. Our paper again offers a direct test to distinguish between the two broad theories of comovement (the price informativeness explanation vs. the price noise explanation). Therefore, we provide substantial evidence based on disagreement among investors from an investor-oriented platform and support firm-specific price movements in which less synchronicity can reflect higher stock price informativeness.

One of the key elements of the disagreement model, as explained by Hong and Stein (2007), is heterogeneous priors; i.e., even if the information is released to all investors simultaneously, each investor will interpret information signals based on their economic model.[72] This assumption is important in our context to understand if investors update their beliefs based on their heterogeneous priors and the arrival of information in financial markets. If so, how does the process of belief formation affect the level of *Disagreement* in financial markets?

To understand this mechanism, we use investors' recommendation revisions on StockTwits. A recommendation revision[73] is defined as the number of times distinct investors on StockTwits update their recommendations from Bullish to Bearish and vice versa for a given stock at time *t* compared to *t-1*. The main intuition behind this idea is that investors will only update their recommendations when they update their economic models based on their priors, thus increasing *Disagreement* among investors and the inflows of firm-specific information in financial markets. We estimate the following equation to understand this mechanism:

$$\begin{aligned} Sync_{i,t} = \beta_0 + \beta_1 Disagreement_{i,t} + \beta_2 Revisions_t + \beta_3\, Disagreement_{i,t} \\ * \; Revisions_t + \beta_4 Revisions_{t-1} + \beta_5 Disagreement_{i,t} \\ * Revisions_{t-1} + \beta_6 Controls + V_i + V_t + V_p + \varepsilon_{i,t} \end{aligned} \quad (7)$$

Where $Revisions_t$ is the total number of recommendation revisions for a given stock at time *t,* and $Revisions_{t-1}$ is the lagged value of recommendation revisions for a given stock at time *t-1.* Recommendation revisions are calculated daily and then aggregated monthly. We include $Revisions_{t-1}$ in Eq. (7) to understand the impact of stale economic models, i.e., the number of revisions in the last month, on *Return Synchronicity.*

[Insert Table 3.*4 here]*

Table 3.4 presents the regression results estimated based on Eq. (7). In Model (1), the coefficient β_2 is negative and significant at the 1% level, suggesting that recommendation revisions provide firm-specific information. The coefficient β_4 is also negative and statistically significant. However, the magnitude is significantly smaller as compared to the coefficient of β_2. These findings suggest that stale economic models can predict the less firm-specific information in financial markets compared to the contemporaneous $Revisions_t$. We conduct a *t*-test on the equality of coefficients of $Revisions_t$ and $Revisions_{t-1}$ and reject the null

[72] Harris and Raviv (1993) and Kandel and Pearson (1995) suggest that differential interpretation of the same signals occurs since investors have different economic models.
[73] It is pertinent to note that Disagreement and Recommendation Revisions are different variables of interest and provide different sets of information. Disagreement is calculated monthly and measures overall disagreement among investors. Recommendation revisions are calculated at the investor level and the number of updates they have made from *t-1* to *t*.

hypothesis that the coefficients are equal. In Models (2) and (3), we examine the interaction between $Revisions_t$ and $Revisions_{t-1}$ with *Disagreement*, respectively, to understand the moderating effect of recommendation revisions on *Return Synchronicity*. As expected, the coefficient of the interaction between *Disagreement* and $Revisions_t$ in Model (2) is negative and statistically significant at the 10% level (–2.5%). These results show that a change in $Revisions_t$ increases the impact of *Disagreement* on *Return Synchronicity,* suggesting that when investors update their economic models, *Disagreement* increases among investors on StockTwits and, consequently, there are higher inflows of firm-specific information in financial markets. The results in Model (3) suggest that $Revisions_{t-1}$ do not affect the impact of *Disagreement* on *Return Synchronicity* since the coefficient of the interaction is not significant. This result also offers useful insights for our analysis by suggesting that *Disagreement* among investors mainly stems from the most recent information.

These findings are consistent with the existing literature. For example, Hong and Stein (2007) argue that investors interpret information based on their economic models even if each investor receives the same information signals, and Banerjee and Kremer (2010) argue that *Disagreement* does not converge since investors update their beliefs due to instantaneous flows of information in financial markets. Therefore, this study provides unique evidence from social media platforms for investors by suggesting how investors on StockTwits updating their beliefs and interactions offers firm-specific information.

3.5.3.2.Media Coverage

Media coverage is an important source of firm-specific information in financial markets. Previous studies have used the aggregated impact of media coverage. For example, Dyck et al. (2008) and Dyck et al. (2013) highlight the role of media coverage in influencing firms' corporate governance and working in favor of special interest groups, respectively. Fang and Peress (2009) argue that media coverage plays a vital role in influencing stock returns in financial markets, and Chahine et al. (2015) present evidence of strategic communication between managers and the media and highlight the role of informative news in financial markets. In our context, the role of *Media Coverage* is twofold. First, higher *Media Coverage* may attract investors to search for further information about the firm and share their opinions/analyses on StockTwits. Second, investors on StockTwits consume any relevant firm-specific information from other channels of information to create discussion threads on StockTwits, as well as updating their recommendations.

In this study, although we are using *Media Coverage* to control the flows of information from traditional information channels, the conspicuous nature of traditional information channels warrants further evidence to investigate the interaction between traditional media and social media (e.g., StockTwits) and how various types of information (e.g., *Breaking News, Full Articles, Press Releases)* may influence investors' opinion on StockTwits. To examine the role of *Media Coverage* in influencing *Disagreement* among investors on StockTwits, we sort firms by *Media Coverage* based on each news type. Specifically, for each news type, *Media Coverage* is aggregated monthly and divided into quintiles, where Q1 is the quintile with no/low *Media Coverage*, and Q5 is the quintile with the maximum *Media Coverage* for that specific news type. Using the regression model estimated in Eq. (5), we run a regression for each news type within a specific quintile. Fig. 3.3 presents the coefficient estimates for our main variable of interest.

In Plot 1, the results from *Overall Media Coverage* show that when moving from Q1 (no/low media coverage) to Q5 (maximum media coverage), the negative impact of *Disagreement* on *Return Synchronicity* tends to increase. The difference between the coefficients of Q5 and Q1 in Plot 1 is ~10%, which is statistically significant at the 1% level. This suggests that *Media Coverage* allows investors to actively engage in discussion threads on StockTwits after interpreting various types of news. In return, investors on StockTwits increase the diffusion of firm-specific information in financial markets. It is pertinent to note that the impact of *Disagreement* does not vanish even if there is no/low media coverage (the coefficient of Q1 is –3.51%), further illustrating the distinct contribution of social media to predicting firm-specific stock-price variation in addition to traditional information channels. These results are consistent with Roll (1988), who presents evidence in favor of low R^2 in the absence of news. He argues that in addition to the release of public information from news, which is capitalized into stock prices, trading activities of informed arbitrageurs also contribute to the capitalization of private information. Overall, these findings are consistent with the existing literature and highlight the key role of media in allowing investors on StockTwits to consume information and increase the velocity of firm-specific information.

[Insert Figure 3.3 Here]

We further estimate regressions after segregating media coverage based on three news types. The first type is *Full Articles* written by authors in the finance and investment industry, the second is *Breaking News*, and the third is *Press Releases* issued by the sample firms. Plots

2 and 3 present the *Disagreement* coefficients of *Full Articles* and *Breaking News,* respectively, for each quintile moving from Q1 (no/low media coverage) to Q5 (maximum media coverage) with a 95% confidence interval. The differences between the coefficients of Q5 and Q1 in Plots 2 and 3 are 10.22% and 11.19%, respectively, which is statistically significant at the 1% level. The economic significance of these results implies that *Full Articles and Breaking News* get considerable attention on StockTwits, thus illustrating the impact of *Disagreement* on *Return Synchronicity* and diffusing the flows of firm-specific information in financial markets at a higher rate.

Plot 4 presents the *Disagreement* coefficient for *Press Releases* issued by the firms for each quintile moving from Q1 (no/low media coverage) to Q5 (maximum media coverage) with a 95% confidence interval. These estimates imply that, although statistically significant, *Press Releases* have the least economic significance compared to *Full Articles* and *Breaking News* in moderating the relationship between *Disagreement* and *Return Synchronicity.* The difference between the coefficients for Q5 and Q1 in Plot 3 is 7.35%, at the 1% level of significance, less than those for *Full Articles* and *Breaking News*. Considering the precise and short text nature of *Press Releases*, it may be challenging for investors to interpret information shared via *Press Releases* unless the commentary is already available in the market. Nekrasov et al. (2019) argue that *Press Releases* are a less common tool for engaging with social media audiences by firms unless the same is shared via the firm's social media handles.

We also estimate regressions after segregating media coverage based on *News Topics* and *News Sources*. Our results remain consistent and further highlight how social media and traditional media complement each other to incorporate firm-specific information in financial markets.

3.5.4. Addressing Endogeneity and Selection Bias

We use two-stage least square (2SLS) regression to address endogeneity concerns and the two-stage Heckman selection model to address self-selection bias.

3.5.4.1.Two-Stage Least Square Regression

Endogeneity is an important concern in our results for multiple reasons. The recent literature investigating the role of media in financial markets has highlighted the role of selective media coverage (Fang & Peress, 2009), sensationalizing rumors (Ahern & Sosyura, 2015), and the impact of macro news (Sheng, 2019). Similarly, Bhagwat and Burch (2016) present strategic tweeting by firms around their earnings announcements, and Clarke et al. (2020) explore how fake news on social media can influence investors' attention. We use a two-stage least square (2SLS) instrumental variable approach to deal with the endogeneity issue. The instrumental variable approach relies on two main assumptions. First, there should be an independent distribution of the excluded instruments' standard errors, and second, the excluded instruments are highly correlated with the endogenous regressors.

Ivković and Weisbenner (2005) present evidence that investors who invest within 250 miles of their geographic proximity earn 3.2% additional annual returns as compared to their nonlocal investments and suggest that such local bias is information-driven. Similarly, Bodnaruk (2009) highlights the role of the local information effect and presents evidence that diversified investors have better access and expertise to process local information. Consequently, such investors earn higher risk-adjusted returns. Therefore, our first instrument variable is *Proximity,* which captures the social distances between investors and the firms' headquarters for whom they have been discussing and sharing ideas on StockTwits. We use *Proximity* as an instrument variable because it satisfies the criteria of a good instrument; *Proximity* is highly and positively correlated with disagreement as it can be an important source of information due to their close proximity to the firm's headquarters. However, the instrument is unlikely to be correlated with the error term in the second stage regression because it is doubtful that the company's stock performance can directly affect *Proximity*. Our first instrument variable is in line with previous studies by Baloria and Heese (2018), who argue that local newspapers are biased towards firms in close proximity, and Peress (2014), who argues that national newspaper strikes affect the flows of information in financial markets, resulting in a 12% lower trading volume on strike days.

Lee and Mas (2012) argue that there are emerging trends of private unionization and highlight the role of labor unions in the financial markets. They argue that financial markets are slow to react to union actions despite their decremental impact on the firms' equity in the long run. Moreover, Wood and Pasquier (2018) provide evidence that social media play a pivotal role in gaining momentum for labor union activities. Hence, such investor-orientated platforms facilitate workers to share their opinion and gain collective identity. Therefore, we use *Labor_Issues* as an additional instrument for disagreement. *Labor_Issues* is defined as the total number of issues related to firms' labor unions aggregated monthly. Our main intuition is that such *Labor_Issues* is positively associated with disagreement as it can exacerbate the number of discussions on social media platforms. However, it is unlikely that the company's stock performance can directly affect *Labor_Issues*. To ensure our instrument variable's economic significance, we only account for labor union strikes, settlements, and layoffs. Overall, we manually collect 553 *Labor_Issues*. To further verify this instrument variable, we cross-check all these issues using news articles from Factiva. Selected newspaper articles discussing *Labor_Issues* are presented in Appendix 3.4.

[Insert Table 3.5 here]

Table 3.5 presents the results from 2SLS regression. The second-stage regression shows the negative association between disagreement and returns synchronicity. The results from the first-stage regression present the positive association between disagreement and two instrument variables. These results are in line with our previous findings, suggesting that the disagreement among investors provides an inflow of firm-specific information. We also check the validity of our instruments based on the following tests. First, the Sargan test is a test for overidentifying restrictions, which tests for the exclusion condition. A *P-value* of the Sargan test, which is higher than 5%, suggests that the excluded instruments are correctly excluded from the estimated equation. For the relevance condition, Stock and Yogo (2005) argue that the weak instruments provide biased instrumental estimators. A rule of thumb for the F-statistic associated with the first stage regression is that it should be greater than 10 (Bound et al., 1995; Staiger & Stock, 1994). The high value of F-statistics (174.11) suggests the higher explanatory power of our instruments, and our two instruments are sufficiently strong to justify inference from the results. Finally, the additional results from the Anderson-Rubin test and the Kleibergen-Paap test reject the null hypothesis (*P-values* smaller than 0.05), suggesting that the model is identified. Therefore, the association between endogenous regressors and the instrument variables are adequate to identify the equation.

3.5.4.2.Two-Stage Heckman Selection Model

Disagreement on StockTwits is a choice for investors, depending on several exogenous factors.[74] However, it is also pertinent to mention that such differences can also be attributed to endogenous factors, since investors on StockTwits can see each other's comments, and popular ideas based on the number of likes and comments can get more attention than the rest. Therefore, such choices can also be determined endogenously. Under such circumstances, self-selection bias could be a potential issue that may influence OLS estimates (Heckman, 1979).

[74] For example, investors' education, investment type, background, and willingness to participate in different types of communication (e.g., like, post, share, etc.).

To address these concerns, Heckman proposes a two-stage model. The first stage is the selection phase, and the second stage is the outcome phase.

We implement the two-stage Heckman (1979) selection model by creating a binary variable to run a probit regression in the selection phase. Our binary variable is the choice between *Disagreement* and *Agreement*. To further ensure the robustness of the model in the first phase, adding an intervening variable that is part of the first stage and not included in the second stage of the model is recommended (Kai & Prabhala, 2007). Specifically, this variable should influence our binary variable in the first stage only. We use the *Proximity* as an additional variable since investors' geographic backgrounds can greatly influence *Disagreement* and Agreement choice. For example, Baik et al. (2016) argue that local Twitter activity can predict higher trading volumes and local social media activity suggest the inflows of private information. In addition to Proximity, we also use the firm-level *R&D/Sales* ratio. This is because firms that allocate more budget to their R&D receive more media coverage. However, it is pertinent to mention that excluding the firm-level *R&D/Sales* ratio does not affect our results in the model's first stage.

Table 3.5, Model (3) presents the probit regression results for the Heckman selection model's first stage. Our main variable of interest is the association of two additional variables with our binary variable. The regression results in Model (3) present that *Proximity* and firm-level *R&D/Sales* ratio can predict *Disagreement,* and it is statistically significant at the 1% level. That is, investors' close proximity to firms' headquarters and firm-level R&D spending can exacerbate the level of *Disagreement* among investors on StockTwits.

From the first stage of the Heckman selection model, we construct an inverse Mills ratio (λ) as an additional regressor to control for self-selection bias in the second phase of the model. Model (4) presents the regression results for the second phase. We find that the Heckman (1979) selection model produces qualitatively similar estimates after correcting for a potential selection problem in our sample. Moreover, the coefficients of λ and *Return Synchronicity* are negative and significant at the 1% level, suggesting that certain observed and unobserved factors can increase the likelihood of a higher level of *Disagreement* among investors on StockTwits, further increasing the flows of firm-specific information in financial markets. For example, suppose one interprets the unobserved component as the *Proximity* of investors. In that case, it can be argued that investors with close social proximity may have information that can influence existing investors' opinions and, consequently, they update their recommendations on StockTwits.

3.5.4.3.Disagreement and Firm Information Environment

The firm information environment plays a pivotal role in influencing information asymmetry in financial markets and reduces the external financing cost for firms with a transparent information environment (Bushman, Piotroski, et al., 2004; Porta et al., 1998). A recent study by Bai et al. (2016) argues that the advances in technology since 1960 have increased price informativeness, and financial markets have become more price efficient. However, Nguyen and Kecskés (2020) argue that technology spillovers increase information asymmetry and the complexity of acquiring information in financial markets. Therefore, the existing literature warrants further evidence to understand the firm information environment's role and its implications for social media platforms for investors. We estimate the following regression to understand the moderating effect of the firm information environment:

$$Sync_{i,t} = \beta_0 + \beta_1 Disagreement_{i,t} + \beta_2 Info\_Proxy + \beta_3 Disagreement_{i,t} * Info\_Proxy + \beta_4 Controls + V_i + V_t + V_p + \varepsilon_{i,t} \tag{8}$$

Where *Info_Proxy* is the set of proxies for the information environment. It is pertinent to note that information asymmetry exists among investors and between firms and investors. For example, Hutton et al. (2009) argue that not all information asymmetry can be associated with exogenous factors. In our context, we are keen to understand the impact of the firm information environment on the level of disagreement among investors on StockTwits. To test this conjecture, we use *Firm Opacity*, *Diversity*, *Industry Concentration,* and *Insider Trading* as different proxies for the firm information environment. The regression results between *Return Synchronicity* and the moderating effect of the firm information environment are presented in Table 3.6.

3.5.4.4.Firm Opacity

Some firms in financial markets are considered opaque as they do not release complete information to financial markets. Lin et al. (2011) argue that firm-information opacity prevents investors from calculating a fair value for the firm and suggest that opaque firms are more likely to face agency issues. Jin and Myers (2006) argue that opaque firms provide less firm-specific information, leaving room for managers to conceal self-serving behaviors. In our case, since disagreement on StockTwits provides firm-specific information, for firms with higher informational opacity, social media platforms for investors, such as StockTwits, can facilitate investors scaling down the further implications of informational opacity by consuming information from StockTwits. We use discretionary accruals (*Disc. Accruals*) as a proxy for *Firm Opacity*. To calculate *Disc. Accruals*, we use Dechow et al. (1995) technique and employ a modified Jones (1991) model.[75] We estimate the following cross-sectional regression based on Fama–French 48 industries for each fiscal year and use residuals to calculate $Disc.\ Accruals_{i,t}$ as follows:

$$Disc.\ Accruals_{it} = \frac{TAC_{it}}{Assets_{it-1}} - \left(\widehat{\lambda_0}\frac{1}{Assets_{it-1}} + \widehat{\lambda_1}\frac{\Delta Revenue_{it} - \Delta Receivables_{it}}{Assets_{it-1}} + \widehat{\lambda_2}\frac{PPE_{it}}{Assets_{it-1}}\right) \tag{9}$$

Model (1) presents the interaction between *Disagreement* and *Info_Proxy (*firm opacity). Our variable of interest is the coefficient of the interaction between *Disagreement* and *Info_Proxy*. This coefficient is statistically significant and negative (–2.40%). These results show that *Firm Opacity* demonstrates the level of *Disagreement* and, consequently, increases the flows of firm-specific information for opaque firms. It is pertinent to note that the stand-alone variable *Info_Proxy* is statistically significant at the 5% level and has a positive coefficient, suggesting that *Firm Opacity* leads to less price-informativeness. These findings suggest that social media platforms for investors such as StockTwits assist investors by offering more insights and analyses to predict stock returns and correctly calculate fair values of opaque firms, which is an otherwise challenging task (Lin et al., 2011).

3.5.4.5.Diversity

Firm-level diversity is defined as the number of business segments and geographic locations in which the firm is operating. Markarian and Parbonetti (2007) argue that diverse firms are complex in nature and face agency issues. This is because diverse firms face challenges in multiple avenues in different geographic locations. Similarly, Bushman, Chen, et al. (2004) present evidence that diversity and the governance structure of firms limit the transparency of firm operations to outsiders. They conclude that there is a clear need to improve the corporate transparency of diverse firms since they are complex in nature.[76]

[75] Total accruals are calculated as income before extraordinary items minus cashflow from operating activities, adjusted for extraordinary items and discontinued operations. Annual data are downloaded from Compustat.
[76] Løwendahl and Revang (1998) highlight the role of the information environment for complex firms and argue that technological changes and sophisticated customers in the post-industrial society have increased firm-level complexity.

Therefore, the flows of firm-specific information for diverse firms can be beneficial for all stakeholders.

In our context, social media platforms for investors can allow stakeholders to access firm-specific information for diverse firms. Following Markarian and Parbonetti (2007), we construct *Diversity* as the natural logarithm of the number of business segments multiplied by the number of geographic segments. Ceteris paribus, the higher the number of business and geographic segments, the higher the uncertainty and demand for information from all stakeholders. Model (2) presents the regression results for *Info_Proxy*. Our main variable of interest is the interaction between *Disagreement* and the *Info_Proxy* variable, i.e., *Diversity*, which is significant at the 1% level with a negative coefficient (–1.8%). This result suggests that disagreement among investors is more pronounced for diverse firms. In contrast, the stand-alone coefficient of Info*_Proxy* (diversity) is insignificant, suggesting that it does not provide any information.

[Insert Table 3.6 *here]*

3.5.4.6.Industry Concentration

Ali et al. (2014) argue that firms in highly concentrated industries disclose less information since the cost of information is higher than the utility of the information. This is because, in more concentrated industries, each firm's market share is comparatively higher than in less concentrated industries. Therefore, firms in these industries may provide reliable information to predict the future demands of industry and market trends. However, industry rivals can use that information to prepare a robust future strategy, resulting in intense market competition, thus increasing the proprietary cost of information disclosure for disclosing firms. Verrecchia (1983) shows that firms with a higher proprietary cost of disclosure disclose less information than firms with a lower proprietary cost of information. The less informative disclosure practices in more concentrated industries warrant further evidence to investigate the role of social media platforms for investors such as StockTwits to assist stakeholders by providing firm-specific information.

To test this argument, we created the *Industry Concentration*[77] measure using the Herfindahl-Hirschman index (HHI). We use firms' total assets for the last three years and two-

[77] Ali et al. (2009) provide substantial evidence that Compustat measures of industry concentration provide mixed results with certain limitations. This is because, when measuring industry concentration, one should also consider the overall industry, which includes private firms. They suggest using US Census measures of industry concentration. In our context, this data set is not available with yearly frequency. Our sample period is only five

digit standard industrial classification (SIC) codes to calculate HHI. Our results remain consistent when we use firms' total sales instead of total assets. Model (3) presents the regression results for the variable *Info_Proxy*. These results show that the coefficient of the interaction between *Disagreement* and *Info_Proxy* is negative and significant at the 1% level (–1.4%), while the variable *Info_Proxy* is insignificant. Our results provide clear evidence, in line with existing literature, that social media platforms for investors assist stakeholders by providing firm-specific information for firms in more concentrated industries.

3.5.4.7.Insider Trading

Insider trading activities can affect the firm information environment since such trades are an important source of private information from firms to market participants. Motivated by Piotroski and Roulstone (2004), we investigate the moderating effect of *Insider Trades* on *Return Synchronicity*. *Insider Trades*[78] are calculated as the absolute value of buy and sell trades scaled by total insider trades in a given month. Model (4) presents the interaction between *Disagreement* and *Info_Proxy*. The coefficient of the interaction is significant at the 1% level and negative (–2.1%), suggesting that higher numbers of insider trades demonstrate the impact of *Disagreement* on *Return Synchronicity* and provide firm-specific information. Given that insiders have an information advantage, disagreement on StockTwits allows stakeholders in financial markets to consume such information and benefit from this advantage, thus reducing information asymmetry under such circumstances.

3.5.5. Disagreement and Salience

An important aspect of limited attention is the conscious allocation of scarce cognitive resources (Kahneman, 1973). Attracting investors' attention in the first place depends on the attention-grabbing characteristics of the stimulus. Such attention-grabbing features are called *Salience* (Fiske & Taylor, 1991). In financial markets, *Salience* can be defined as the information itself (*Information signals*[79]), or the *Sources*[80] of information. Such stimuli with different levels of *Salience* compete in financial markets to attract investors' attention.

years, while the US Census measure of industry concentration data is only available every five years. However, our findings are consistent with previous literature and using alternative proxies for industry concentration, suggesting that this caveat does not affect our results.

[78] Data are collected from Thomson Reuters' insider trading database. We use transaction codes P and S, and role codes CB, CEO, CO, GC, and P.

[79] This includes any material information that can be useful for investors to predict future prices and firms' future earnings.

[80] This includes industry professionals and influencers who analyze financial markets.

However, only stimuli with higher *Salience* can increase the marginal utility of information acquisition for investors wishing to remain in the competition (Hirshleifer et al., 2011; Hirshleifer & Teoh, 2003). Previous studies have mainly discussed the impact of limited attention without explaining the effects of *Salience*. However, it plays a vital role in the attention-allocation process. In recent studies by Li et al. (2019) and Huang et al. (2018), they argue that *Salience* is a key feature, based on which investors determine the quality of information signals in financial markets. Therefore, our next strand of investigation is to understand the role of *Salience*. Specifically, our aim is to understand the impact of *Salience* on disagreement among investors on StockTwits. We divide *Salience* into *Information Signals* and the *Heterogeneity of Investors*.

3.5.5.1.Information Signals

We broadly define the attention-grabbing characteristics of information signals on StockTwits based on their *Network* and *Social Media Attention* (SMA). The *Network* is defined as the reach of information signals in an extensive social network; i.e., the larger the network, the greater the reach of information signals. Similarly, *SMA* is further divided into three subgroups based on the number of ideas (*Ideas*); the *Popularity* of ideas, which is the number of likes each idea receives; and, finally, *Discussion* on StockTwits, which is defined as the number of replies a specific idea has on StockTwits. We use *Salience* as a proxy to represent all these variables.

[Insert Table 3.7 *here]*

Table 3.7 presents the regression results for the *Salience* groups, i.e., *Network* and *SMA*, respectively. One of the distinguishing features of social media is its vast network of users. Because of these extensive social networks, the velocity of information diffusion on social media is far higher than on any other traditional information channel. Model (1) presents the results for the interaction between *Network* and *Disagreement*. Our main variable of interest is the interaction coefficient between *Disagreement* and *Network*, which is negative and significant at the 1% level (–7.8%). The association between *Disagreement* and *Network* further suggests that extensive social networks demonstrate the impact of *Disagreement* on *Return Synchronicity*. The result in Model (1) implies that when influential investors[81] post ideas on StockTwits, this allows others to follow the lead of those ideas across a broader

[81] Any investor with a large number of followers on StockTwits is known as an influential investor.

spectrum, consequently increasing the level of *Disagreement* and prompting a higher inflow of firm-specific information in financial markets.

Models (2) to (4) present the regression results for the interaction between *Disagreement* and SMA subgroups, Ideas, Popularity, and Discussion, respectively. All three coefficients of the interactions between these Salience groups and *Disagreement* are negative and statistically significant at the 1% level. It is pertinent to note that attention allocation and responses to any stimuli are simultaneous actions. For example, investors can allocate attention by posting ideas, liking ideas, or participating in discussion threads on StockTwits. However, our results also suggest that posting ideas is the most popular way of allocating attention on StockTwits by comparing the magnitudes of the coefficients associated with the variable of Salience across SMA subgroups, i.e., Ideas, Popularity, and Discussion, based on both non-interacted and interacted with disagreement.

3.5.5.2.Heterogeneity of Investors

Our final *Salience* group is the *Heterogeneity of Investors* on StockTwits. To examine the impact of the heterogeneity of investors on the level of disagreement among investors, we first define heterogeneity based on the presence of unique investors. Second, we define heterogeneity based on the self-disclosed investors' experience and investment approaches on StockTwits[82]. For example, investors' experience on StockTwits is broadly categorized into three groups: professional, intermediate, and novice. Similarly, investment approaches are categorized into momentum, technical, fundamental, and value[83]. The results are presented in Table 3.8.

To understand the overall impact of the presence of unique investors, we examine the moderating effect of unique investors on *Disagreement* in Model (1). The coefficient of the interaction term is negative and statistically significant at the 1% level (–7.9%), suggesting that the impact of *Disagreement* on *Return Synchronicity* is more pronounced when *Disagreement* is from more diverse investors. This result is consistent with our previous findings, suggesting

[82] Investors disclose such information voluntarily and these disclosures are not a requirement on StockTwits.
[83] In addition to these investment approaches, investors have the choice to select from global macro and growth investment approaches. However, these investment approaches are less popular. In our sample, there are less than 1% investors who choose such investment approaches. Following Cookson and Niessner (2019), we exclude such approaches. However, including investors with such investment approaches does not change our results.

that an increase in heterogeneity means higher salience of information signals is attracting more investors to share their ideas on StockTwits.

The next category for the *Heterogeneity of Investors* is investors' self-disclosed experience on StockTwits. For this purpose, we calculate the within-group disagreement among investors who disclose their investment experience on StockTwits. Ideally, professional investors should take the lead to facilitate the flows of firm-specific information as compared to novice investors, and this is what we find in our regression results presented in Model (2). Overall, the professional investors' coefficient magnitude is the largest, followed by intermediate and novice investors. However, we could not reject the null hypothesis of the equality of coefficients of professional and novice investors using the *t*-test, given that the *P*-value is 0.1347. [84]

[Insert Table 3.8 here]

Previous studies such as Jegadeesh and Titman (1993) highlight the role of momentum investing and suggest that momentum investors can earn higher returns. Similarly, a recent study by Hillert et al. (2014) suggests that stock covered by media has significantly higher momentum. In our case, StockTwits plays a pivotal role in diffusing the information from various information channels. Therefore, it is important to examine the impact of investment approaches on Return Synchronicity. Model (3) presents regression results using within-group disagreement for momentum, technical, fundamental, and value investors. The results show that within-group disagreement among investors decreases Return Synchronicity, indicating higher inflows of firm-specific information. These results remain consistent across all investment approaches. However, the negative impact of *Disagreement* on Return Synchronicity is higher for Momentum (6.2%), followed by Technical (5.1%) and Fundamental (4.4%). The impact of Value *Disagreement* on Return Synchronicity (2.7%) is less than half of that for Momentum. To further understand the difference between coefficients of *Disagreement* for investment approaches, we conduct the *t*-test for the equality of coefficients between momentum and value investors, and we reject the null hypothesis that two coefficients are equal. It is pertinent to note that we cannot reject the null hypothesis when comparing coefficients of disagreement of Momentum and Technical investors. This is because

[84] We follow a conservative approach by using two-dimensional clustering (firm and month). However, when we only use single clustering at the firm level rather than double clustering, the *t*-test of the equality of the coefficients is significant, suggesting that professional investors play an important role in diffusing firm-specific information as compared to novice investors.

Momentum investors may also follow Technical investment approaches such as moving averages.

Overall, our results provide compelling evidence that the *Salience* on StockTwits, which comprises *Information Signals* and the *Heterogeneity of Investors,* plays a crucial role in facilitating investors to efficiently allocate their attention and show the flows of firm-specific information in financial markets. These findings are consistent with recent studies suggesting that social interactions play an essential role in influencing investors' behaviors in financial markets (Hirshleifer, 2019).

3.6. Robustness Checks

3.6.1. Alternative Proxies for Return Synchronicity

In our study, we use Carhart (1997), four-factor model, to derive the value of R^2 and then calculate *Return Synchronicity*. As a further check on the robustness of our model, we also use alternative proxies for *Return Synchronicity*. In this vein, Morck et al. (2000) model (henceforth MYY) offers a slight variation to derive the value of R^2 as follows:

$$\begin{aligned} Ret_{i,t} = \lambda_0 + \lambda_1 Market\ Return_{i,t} + \lambda_2 Industry\ Return_{i,t} + \lambda_3 Market\ Return_{i,t-1} \\ + \lambda_4 Industry\ Return_{i,t-1} + \varepsilon_{i,t} \end{aligned} \tag{10}$$

Where $Industry\ Return_{i,t}$ is calculated based on two-digit SIC industry codes. In other studies, Peng and Xiong (2006) and Antón and Polk (2014) measure *Return Synchronicity* as a times series of Pearson's correlation coefficients between the firm and market return:

$$CORR_{k,m} = \frac{COV(R_k, R_m)}{\sigma_{R_k}, \sigma_{R_k}} \tag{11}$$

In addition to using alternative proxies, we extend the four-factor and MYY models by using the value of $Adj.R^2$ instead of R^2. This is a conservative approach since a large chunk of sample observations are lost because $Aj.R^2$ cannot be calculated with fewer observations. Our results from the robustness tests using these alternative proxies for *Return Synchronicity* are presented in Table 3.9. These results are consistent with our model, and our main variable of interest, *Disagreement,* is significant at the 1% level with a negative coefficient, even after using alternative proxies for *Return Synchronicity*.

[Insert Table 3.*9 here]*

3.6.2. Alternative Machine-Learning Approaches for Recommendation Predictions

Machine-learning prediction models are selected based on the quality of their predictions and the availability of parameters that match the data set. In our study, we use the Random Forest Decision Trees (*RFDT*) approach. As a further check on the robustness of our model and to understand if variations in the recommendation predictions model might affect our results, we also predict recommendations using Support Vector Machine (*SVM*) and Maximum Entropy (*MaxEnt*). It is pertinent to note that in all three prediction models, our training data set remains the same, and we use tenfold cross-validation and feature selection. The accuracy and F1 score for SVM are 80% and 87%, respectively. Similarly, the accuracy and F1 score for MaxEnt are 74% and 83%, respectively. The results presented in Table 10 complement our findings, suggesting that when using alternative prediction models, our results remain consistent.

[Insert Table 3.*10 here]*

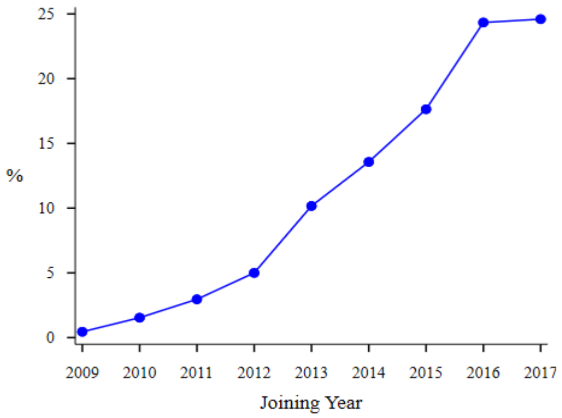
3.7. Chapter Conclusion

This paper provides new evidence that social media platforms for investors, such as StockTwits, assist investors by providing firm-specific information. More importantly, our study complements the existing literature by offering substantial evidence that social media platforms for investors provide firm-specific information that can help investors make their investment decisions. These findings are consistent with Hong and Stein (2007) heterogenous-agent framework model and existing literature on behavioral finance. Our analysis is based on more than 12 million posts from StockTwits posted by 162,836 unique investors from January 2013 to December 2017, discussing 956 US-listed companies. To predict investors' recommendations and measure the level of disagreement among investors on StockTwits, we use Random Forest Decision Trees as our main prediction model. To the best of our knowledge, this is the first study to highlight the role of discussions on social media in providing firm-specific information.

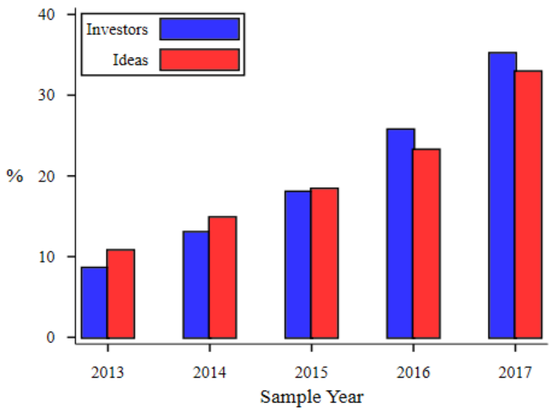
This study contributes to the existing literature on the role of social media in financial markets in the following areas. First, discussions on StockTwits result in higher inflows of firm-specific information. Second, overall media coverage and investors' recommendation revisions contribute to investors' talk on StockTwits, consequently increasing disagreement and prompting higher inflows of firm-specific information. We extend our investigation to assess the effect of firms' information environment. We find that disagreement among investors on StockTwits plays a pivotal role in the supply of firm-specific information for firms with higher Opacity, Diversity, Industry Concentration, and Insider Trades. Third, using the salience of information signals on StockTwits, we show how investors on social media platforms follow leads from influencers, resulting in new investors opting to participate in discussions. Consequently, such interactions result in higher levels of disagreement and information diffusion in financial markets. These findings are robust to endogeneity, sample selection bias, and using alternative measures of return synchronicity and recommendation predictions.

Our findings have practical implications for portfolio managers and investors. Portfolio managers can develop multiple portfolio strategies based on the level of disagreement and the salience of information signals on social media platforms for investors. After considering the practical implications of investors' opinions on StockTwits, Thomson Reuters Eikon and Bloomberg Terminals have already embedded an online version of StockTwits in their platforms to assist investors and portfolio managers. Unlike other social media platforms for investors, access to StockTwits is free. This motivates investors to sift through any information based on cashtags, investment philosophies, and investment approaches, as well as the period of investment. Our findings complement the emerging literature in social finance by suggesting that the economic significance of social interactions on social media platforms for investors plays a pivotal role in investment decision-making. Finally, this study also contributes to the emerging literature on the role of big data and machine learning in finance by using one of the largest training data sets for recommendation predictions and comparing prediction models to validate the robustness of our results.

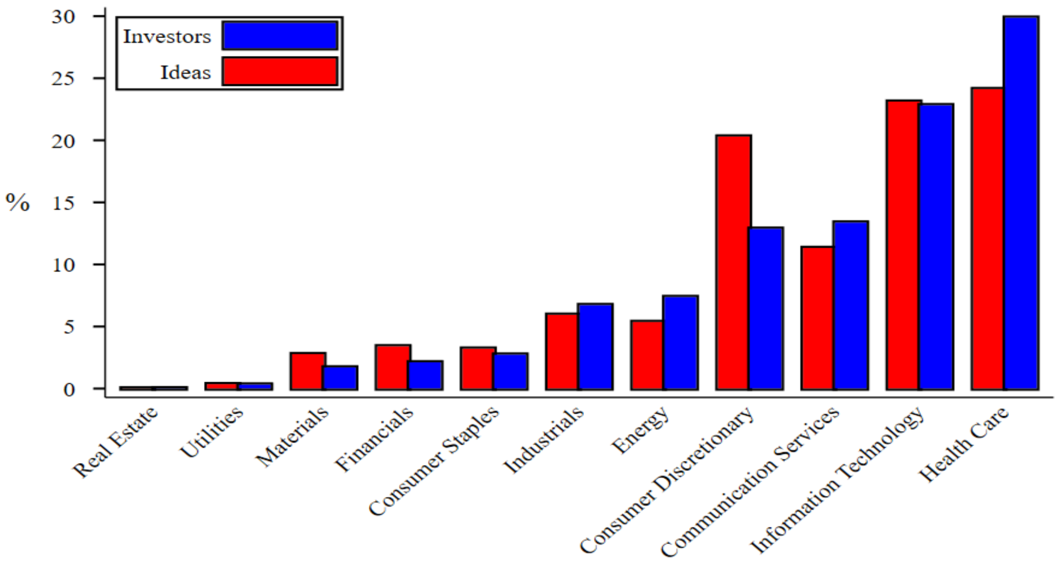
Figure 3.1: Distribution of Investors and Ideas on StockTwits



Panel A: Distribution of Investors



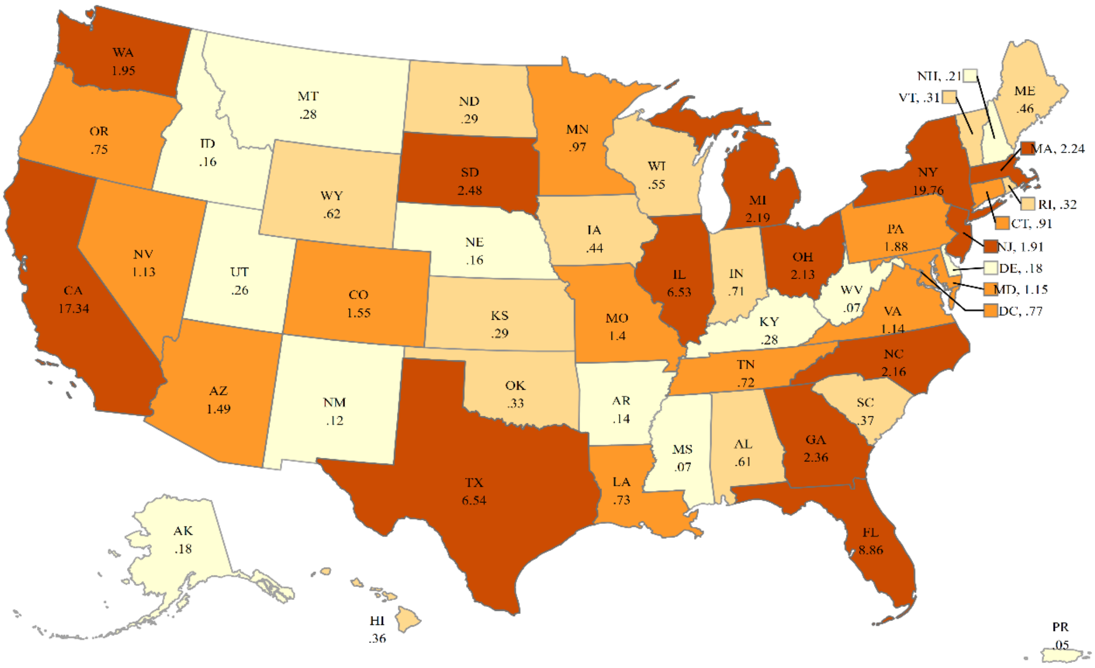
Panel B: Distribution of Ideas and Investors



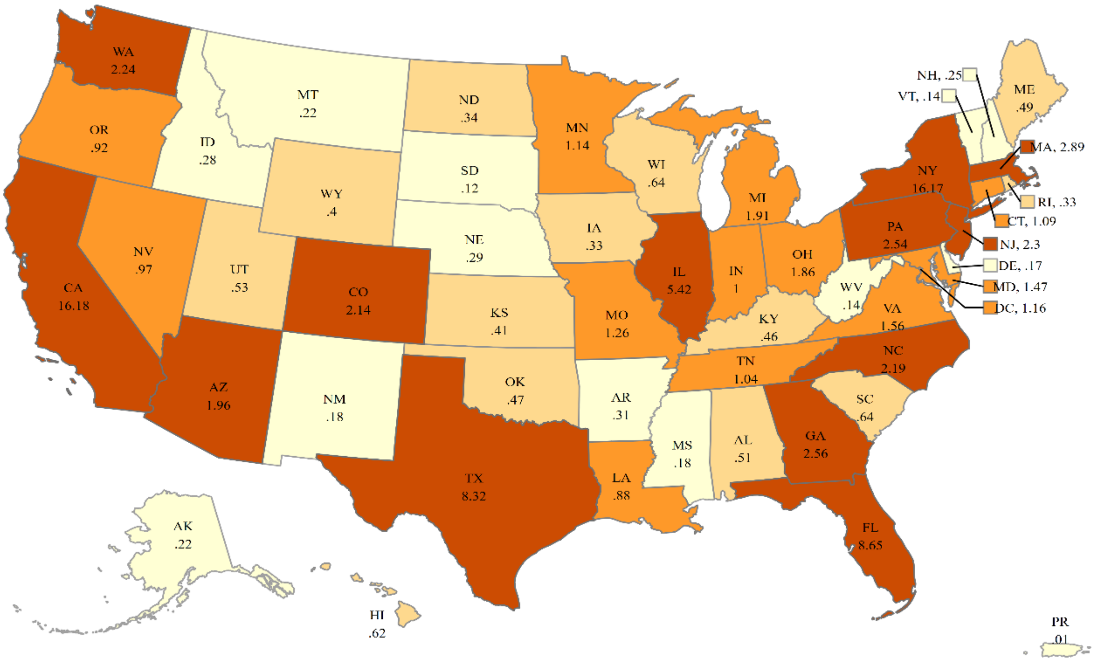
Panel C: Distribution of Ideas and Investors across GICS Sectors

Notes: Fig. 3.1 presents the distribution of ideas and investors in the StockTwits sample. Panel A presents the distribution of investors based on the year they joined StockTwits. Panel B presents the distribution of investors and ideas posted by these investors on StockTwits during the sample period. Panel C presents the distribution of investors and ideas based on the Global Industrial Classification System (GICS) sectors.

Figure 3.2: Geographic distribution of StockTwits' ideas and investors across the USA



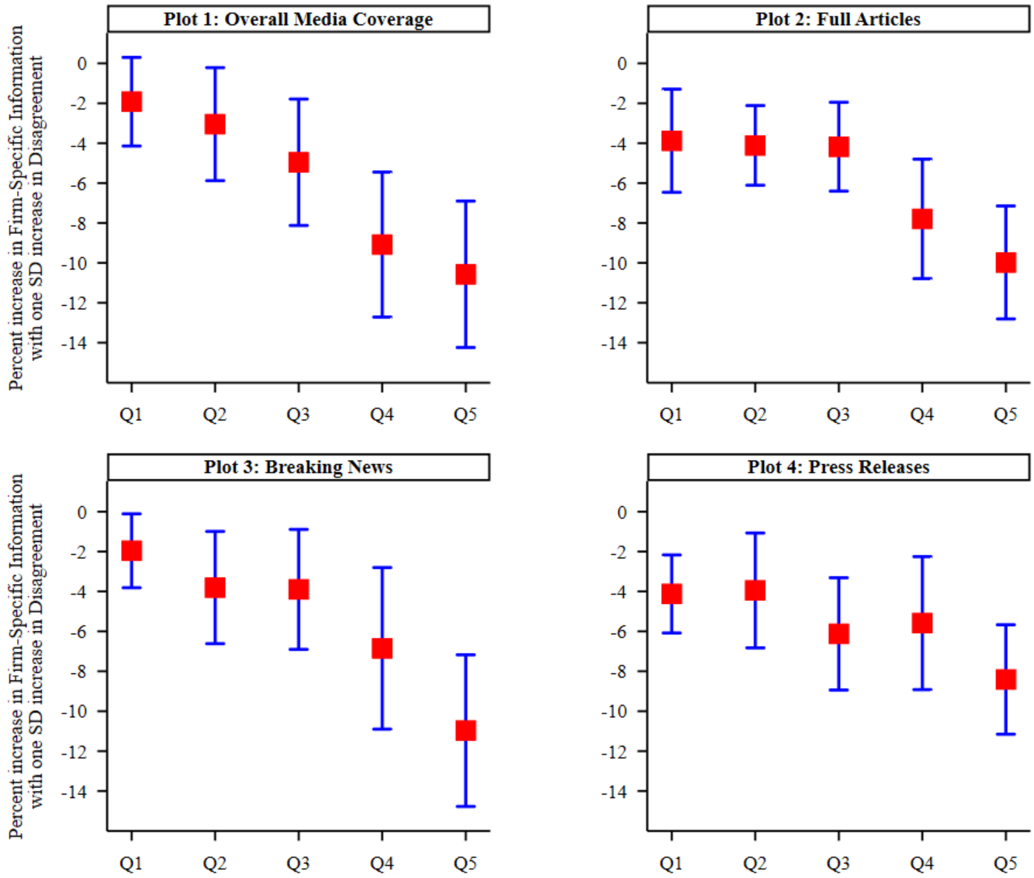
Panel A: Distribution of Ideas



Panel B: Distribution of Investors

Notes: Fig. 3.2 presents the distribution of StockTwits' ideas and investors across the USA, respectively. Users' locations are collected from their StockTwits public profiles. Users' locations are further cleaned using text analysis and then matched with US state-level coordinates obtained from US Census website geographic data.

Figure 3.3: Disagreement & Media Coverage



Notes: Fig. 3.3 presents coefficient estimates from the multivariate fixed-effect regression between *Disagreement* and *Return Synchronicity* along with their corresponding 95% confidence intervals. The dependent variable is Return Synchronicity, which is calculated using Carhart's (1997) four-factor model. Disagreement is derived from users' recommendations on StockTwits. Media coverage is segregated into groups based on the type of information. Each group is divided into quintiles based on the number of news articles published monthly, where quintile 1 presents firm-month observations with no/low media coverage, and quintile 5 presents firm-month observations with the highest media coverage. The standard regression model in equation (5) is used for each quintile after excluding media coverage as an explanatory variable. The type groups are defined as the number of press releases issued by the sample firm, full articles as the number of detailed articles that discuss the sample firms, and breaking news as the number of news flashes that explicitly mention the sample firm. The regressions are estimated using time, firm, and industry fixed-effects. The sample consists of 956 firms with 52,888 firm-month observations for the sample period of 2013–2017. Variable definitions are presented in Appendix 3.2.

Table 3.1: Summary Statistics

Notes: The table reports the summary statistics of StockTwits' ideas and investor-level information, media coverage, and firm-level characteristics of the sample firms. All the variables are defined in Appendix 3.2. Disagreement and Revisions are calculated using investors' recommendations on StockTwits. These recommendations are predicted using Random Forest Decision Trees (RFDT) method. Further details about this method are discussed in Appendix 3.1.

Panel A: StockTwits Ideas and Investor Level Information

	Mean	SD	P10	P25	Median	P75	MAX
Disagreement	0.53	0.26	0	0.43	0.59	0.7	1
Revisions	3.61	2.87	0	1.39	3.22	5.21	17.27
Network	12.97	2.18	11.08	12.35	13.28	14.13	18.27
Investors	3.75	1.30	2.20	2.94	3.69	4.49	9.55
Ideas	3.99	1.44	2.30	3.09	3.89	4.78	10.81
Popularity	2.06	1.98	0	0	1.61	3.22	10.90
Discussion	2.60	2.34	0	0	2.30	4.26	12.36
Proximity	0.64	1	0	0	0	1.10	7.79

Panel B: Media Coverage

	Mean	SD	P10	P25	Median	P75	MAX
Media Coverage (Overall)	2.95	1.06	1.79	2.3	2.94	3.56	8.55
Breaking News	1.61	1.14	0	0.69	1.61	2.4	6.07
Full Articles	1.85	1.22	0	1.10	1.79	2.64	7.44
Press Release	1.61	1.17	0	0.69	1.61	2.4	6.31

Panel C: Firm-Level Characteristics

	Mean	SD	P10	P25	Median	P75	MAX
Return Synchronicity	-0.37	1.04	-1.70	-1.02	-0.33	0.33	3.85
Firm Size	19.33	51.57	0.06	0.59	3.21	15.48	867.51
Analyst Coverage	12.07	8.29	2	5	11	17	54
Leverage	0.25	0.25	0	0.04	0.21	0.37	3.44
Market/Book Ratio	4.32	6.37	0.41	1.36	2.63	4.92	110.53
Adv/Sales	0.02	0.22	0	0	0	0.02	11.36
Sales Growth	0.60	0.42	0	0.49	0.69	0.75	6.83
ROA	0.04	0.30	-0.33	0.02	0.10	0.17	1.42
Earnings Volatility	0.07	0.14	0.01	0.01	0.03	0.08	5.19
Real GDP	0.02	0.01	0.01	0.02	0.02	0.03	0.04
Firm Opacity	-0.24	0.28	-0.59	-0.39	-0.18	-0.06	2.16
Diversity	1.66	1.15	0	0.69	1.95	2.71	4.29
Competition	0.01	0.02	0	0	0	0.01	0.35
Insider Trading	0.45	0.49	0	0	0	1	1

Table 3.2: Return Synchronicity and Disagreement

Notes: The table reports the regression results of Disagreement and Return Synchronicity. The dependent variable is Return Synchronicity, which is calculated using the Carhart (1997) four-factor model. Disagreement is derived from the investors' recommendations on StockTwits. All right-hand side variables are standardized. The sample consists of 956 firms with 53,778 firm-month observations for the sample period between 2013 - 2017. The regressions are estimated using the ordinary least square (OLS) method with time, firm, and industry fixed effects in Model (1) – (3). However, in Model (4) and (5), regressions are estimated using OLS and Fama McBeth regression without fixed effects, respectively. Variable definitions are presented in appendix 3.2. *, **, and *** represent significance at 10%, 5%, and 1%, respectively. The standard errors are reported in brackets and are robust to heteroscedasticity and clustered at the firm and month levels in Model (1) – (4). In Model (5), standard errors are adjusted for cross-sectional dependence based on Fama-McBeth regression.

	OLS				**Fama-MacBeth**
	1	**2**	**3**	**4**	**5**
Disagreement	-0.069*** [0.0119]	-0.054*** [0.0117]	-0.057*** [0.0109]	-0.079*** [0.0120]	-0.075*** [0.0064]
Media Coverage		-0.221*** [0.0266]	-0.227*** [0.0237]	-0.000 [0.0190]	-0.050*** [0.0104]
Analyst Coverage			0.116*** [0.0145]	0.256*** [0.0168]	0.199*** [0.0144]
Leverage			-0.095*** [0.0238]	-0.027** [0.0108]	-0.022*** [0.0064]
Adv/Sales			-0.015* [0.0071]	-0.030*** [0.0096]	-0.081*** [0.0146]
Market/Book Ratio			-0.032** [0.0112]	-0.026** [0.0109]	-0.027*** [0.0085]
Firm Size			0.203*** [0.0578]	0.096*** [0.0249]	0.281*** [0.0580]
ROA			0.053*** [0.0160]	0.017 [0.0150]	-0.009 [0.0097]
Earnings Volatility			0.015** [0.0055]	-0.033* [0.0162]	-0.052*** [0.0083]
Sales Growth			0.052 [0.0302]	0.099*** [0.0280]	0.086*** [0.0148]
Real GDP_{t-1}			-0.075* [0.0410]	-0.072 [0.0412]	-0.001 [0.0506]
Fixed Effects	Y	Y	Y	N	N
Adj/Avg. R-squared	0.195	0.211	0.228	0.089	0.128/0.138
Firms	956	956	956	956	956
Observations	53,778	53,778	52,888	52,888	52,888

Table 3.3: Disagreement and Price Informativeness

Notes: The table reports the regression results of the Disagreement and the moderating effect of price informativeness. The dependent variable is cumulative abnormal returns (CAR) calculated quarterly. The benchmark return is calculated based on 5 x 5 Size and Book to Market portfolios to calculate CAR. Change in earnings (ΔEarningst) is used as a proxy of price informativeness and is calculated as firms' earnings at time t minus firms' earnings at time t-1 divided by the beginning of the time t market value equity of the firm. Disagreement is derived from the investors' recommendations on StockTwits and calculated at a quarterly frequency. The sample consists of 949 firms with 17,309 firm-quarter observations for the sample period between 2013 - 2017. The regressions are estimated using the ordinary least square (OLS) method with firm fixed effects in Model (1) – (3). However, in Model (4) - (6) and (7) - (9), regressions are estimated using OLS and Fama-McBeth regression without fixed effects, respectively. *, **, and *** represent significance at 10%, 5%, and 1%, respectively. The standard errors are reported in brackets and are robust to heteroscedasticity and clustered at the firm levels in Model (1) – (6). In Model (7) – (9), standard errors are adjusted for cross-sectional dependence based on Fama-McBeth regression.

	OLS						Fama-MacBeth		
	1	2	3	4	5	6	7	8	9
Disagreement		-0.196*** [0.0119]	-0.185*** [0.0114]		-0.164*** [0.0107]	-0.190*** [0.0104]		0.360 [0.5211]	0.170 [0.4483]
$\Delta Earnings_{t+1}$	-0.001 [0.0015]	-0.016*** [0.0054]	-0.015*** [0.0058]	-0.002 [0.0014]	-0.012** [0.0058]	-0.012** [0.0057]	0.120** [0.0499]	-0.111 [0.1123]	-0.118 [0.1091]
*Disagreement * $\Delta Earnings_{t+1}$*		0.021*** [0.0070]	0.022*** [0.0070]		0.015* [0.0077]	0.015** [0.0074]		0.385* [0.1924]	0.390** [0.1839]
$\Delta Earnings_{t}$	0.006*** [0.0019]	0.008 [0.0048]	0.005 [0.0047]	0.006*** [0.0021]	0.009* [0.0049]	0.008 [0.0049]	0.030 [0.0378]	-0.001 [0.1761]	-0.049 [0.1650]
*Disagreement * $\Delta Earnings_{t}$*		-0.004 [0.0081]	-0.002 [0.0079]		-0.005 [0.0090]	-0.003 [0.0087]		0.030 [0.3055]	0.098 [0.2852]
$\Delta Earnings_{t-1}$	-0.002 [0.0018]	0.011*** [0.0040]	0.007* [0.0038]	-0.002 [0.0017]	0.008* [0.0044]	0.007 [0.0043]	0.058** [0.0214]	0.150 [0.1067]	0.120 [0.1001]
*Disagreement * $\Delta Earnings_{t-1}$*		-0.023*** [0.0056]	-0.018*** [0.0053]		-0.017*** [0.0065]	-0.015** [0.0061]		-0.153 [0.1749]	-0.116 [0.1619]
Controls	N	N	Y	N	N	Y	N	N	Y
Fixed Effects	Y	Y	Y	N	N	N	N	N	N
Adj/Avg. R-squared	0.036	0.057	0.095	0.004	0.022	0.049	0.012/0.014	0.044/0.051	0.083/0.091
Firms	946	946	946	949	949	949	949	949	949
Observations	17,306	17,306	17,306	17,309	17,309	17,309	17,309	17,309	17,309

Table 3.4: Return Synchronicity, Disagreement, and Revisions

Notes: The table reports the regression results of Disagreement and Recommendation Revisions of investors on StockTwits. The dependent variable is Return Synchronicity, which is calculated using the Carhart (1997) four-factor model. Disagreement is derived from the investors' recommendations on StockTwits. Recommendation revisions are calculated as the number of revisions from Bullish-Bearish and vice versa by each investor at time t to t-1 and aggregated at a monthly frequency. All right-hand side variables are standardized. The sample consists of 956 firms with 53,778 firm-month observations for the sample period between 2013 - 2017. The regressions are estimated using time, firm, and industry fixed effects. Variable definitions are presented in appendix 3.2. *, **, and *** represent significance at 10%, 5%, and 1%, respectively. The standard errors are reported in brackets and are robust to heteroscedasticity and clustered at the firm and month levels.

	1	**2**	**3**
Disagreement		0.010 [0.0114]	-0.051*** [0.0094]
Revisions	-0.217*** [0.0212]	-0.230*** [0.0229]	
*Disagreement * Revisions$_t$*		-0.025* [0.0118]	
Revisions$_{t-1}$	-0.020** [0.0084]		-0.025** [0.0084]
*Disagreement * Revisions$_{t-1}$*			-0.003 [0.0036]
Media Coverage	-0.138*** [0.0226]	-0.141*** [0.0227]	-0.218*** [0.0239]
Analyst Coverage	0.113*** [0.0151]	0.112*** [0.0146]	0.115*** [0.0146]
Leverage	-0.037 [0.0232]	-0.038 [0.0232]	-0.092*** [0.0233]
Adv/Sales	-0.014** [0.0064]	-0.013* [0.0070]	-0.016** [0.0064]
Market/Book Ratio	-0.044*** [0.0112]	-0.039*** [0.0109]	-0.038*** [0.0111]
Firm Size	0.168** [0.0560]	0.173** [0.0559]	0.193*** [0.0574]
ROA	0.057*** [0.0148]	0.054*** [0.0147]	0.056*** [0.0159]
Earnings Volatility	0.015** [0.0051]	0.015** [0.0051]	0.016** [0.0060]
Sales Growth	0.076** [0.0322]	0.070* [0.0321]	0.057* [0.0297]
Real GDP$_{t-1}$	-0.073* [0.0392]	-0.068 [0.0398]	-0.082* [0.0401]
Fixed Effects	Y	Y	Y
Adj. R-squared	0.249	0.248	0.230
Firms	956	956	956
Observations	52,888	52,888	52,888

Table 3.5: Testing for Endogeneity and Selection Bias

Notes: The table reports the results from two-stage least square (2SLS) regression using the instrumental variable approach and two-stage Heckman Selection model. There are two instruments in the 2SLS regression. Proximity is defined as the number of investors who post ideas on StockTwits while discussing the sample firms and who have the same US state where the firms' headquarter is located. Labor_Issues, defined as the total number of issues related to firms' labor unions aggregated monthly. Sanderson-Windmeijer F test of excluded instruments is presented as S-W F-statistics, Kleibergen-Paap Wald F statistic is presented as K-P Wald F-statistics, and Anderson-Rubin Wald test is presented as A-R Wald F-statistics. In the two-stage Heckman selection model, the first stage selection equation is estimated by probit regression, where the dependent variable is 1 in case of Disagreement and 0 otherwise. In the second stage, the dependent variable is the Return Synchronicity, and the inverse mills ratio (λ) adjusts for the nonzero mean of error terms. All right-hand side variables are standardized. The sample consists of 956 firms with 52,888 firm-month observations for the sample period between 2013 - 2017. Variable definitions are presented in appendix 3.2. *, **, and *** represent significance at 10%, 5%, and 1%, respectively. The standard errors are reported in brackets and are robust to heteroscedasticity and clustered at the firm level.

	2SLS Regression		**Heckman Selection**	
	1	**2**	**3**	**4**
	First Stage	**Second Stage**	**Selection**	**Outcome**
Disagreement		-0.721*** [0.066]		-0.029*** [0.008]
Proximity	0.168*** [0.009]		0.579*** [0.013]	
Labour_Issues	0.009*** [0.002]			
R&D/Sales			0.251*** [0.097]	
Media Coverage	0.126*** [0.008]	-0.120*** [0.014]		-0.031*** [0.006]
Firm Size	0.011 [0.014]	0.204*** [0.020]	-0.011 [0.007]	0.566*** [0.017]
Analyst Coverage	0.070*** [0.014]	0.162*** [0.016]	0.214*** [0.007]	
Leverage	0.131*** [0.020]	0.004 [0.025]		-0.009* [0.005]
Market/Book Ratio	-0.066*** [0.013]	-0.077*** [0.014]		0.035*** [0.005]
Adv/Sales	0.000 [0.004]	-0.014** [0.006]	-0.002 [0.009]	-0.019*** [0.004]
Sales Growth	0.105*** [0.027]	0.126*** [0.036]	0.216*** [0.018]	0.066*** [0.012]
ROA	0.005 [0.013]	0.058*** [0.017]	0.017** [0.009]	0.046*** [0.006]
Earnings Volatility	-0.006 [0.008]	0.011 [0.008]	0.049*** [0.009]	-0.029*** [0.006]
Real GDP_{t-1}	0.008 [0.005]	-0.066*** [0.006]		-0.071*** [0.005]
λ				-0.098*** [0.033]
Fixed Effects	Y	Y	N	N
Firms	956	956	956	956
Observations	52,888	52,888	53,221	52,888
S-W F-statistics	174.11***			
K–P Wald F-statistic	174.11***			
A-R Wald F-statistics	86.68***			
Sargan *P-Value*		0.345		

Table 3.6: Disagreement and Firm information environment

Notes: The table reports the regression results of the Disagreement and the moderating effect of the firm's information environment. Variable Info_Proxy is the proxy of variables from the firm information environment. The dependent variable is Return Synchronicity, which is calculated using the Carhart (1997) four-factor model. Disagreement is derived from the investors' recommendations on StockTwits. Firm opacity is calculated as a measure of firms' accrual quality based on the extended Jones (1991) model. Diversity is the natural logarithm of the number of business segments of the firm multiplied by the number of geographic segments. Competition is based on firm-level assets calculated as a proxy of industry competition using the Herfindahl-Hirschman Index. Insider trading is calculated as the absolute value of the difference between buying and selling insider trades scaled by the total insider trades in a given month of a sample firm. All right-hand side variables are standardized. The sample consists of 956 firms with 49,723 firm-month observations in the Model (1) and 52,888 firm-month observations in the Model (2) – (4) for the sample period between 2013 - 2017. The regressions are estimated using time, firm, and industry fixed effects. Variable definitions are presented in appendix 3.2. *, **, and *** represent significance at 10%, 5%, and 1%, respectively. The standard errors are reported in brackets and are robust to heteroscedasticity and clustered at the firm and month levels.

	1	**2**	**3**	**4**
	Firm Opacity	**Diversity**	**Ind. Concentration**	**Insider Trading**
Disagreement	-0.052*** [0.0074]	-0.047*** [0.0071]	-0.046*** [0.0075]	-0.058*** [0.0106]
Info_Proxy	0.035** [0.0128]	-0.013 [0.0193]	-0.160 [0.0904]	0.017* [0.0087]
*Disagreement * Info_Proxy*	-0.024*** [0.0050]	-0.018*** [0.0055]	-0.014*** [0.0044]	-0.021*** [0.0053]
Media Coverage	-0.232*** [0.0201]	-0.233*** [0.0199]	-0.234*** [0.0204]	-0.234*** [0.0247]
Analyst Coverage	0.100*** [0.0125]	0.099*** [0.0096]	0.106*** [0.0122]	0.115*** [0.0144]
Leverage	-0.056** [0.0226]	-0.061** [0.0214]	-0.055** [0.0218]	-0.094*** [0.0238]
Adv/Sales	-0.015** [0.0064]	-0.016** [0.0060]	-0.015** [0.0060]	-0.015* [0.0070]
Market/Book Ratio	-0.006 [0.0077]	-0.010 [0.0072]	-0.004 [0.0075]	-0.033** [0.0110]
Firm Size	0.046* [0.0240]	0.048* [0.0238]	0.049* [0.0243]	0.205*** [0.0576]
ROA	0.044** [0.0152]	0.050*** [0.0149]	0.053*** [0.0153]	0.053*** [0.0160]
Earnings Volatility	0.021** [0.0091]	0.022** [0.0088]	0.022** [0.0088]	0.015** [0.0055]
Sales Growth	0.031 [0.0208]	0.044* [0.0211]	0.044* [0.0209]	0.052 [0.0301]
Real GDP_{t-1}	-0.144** [0.0595]	-0.146** [0.0590]	-0.147** [0.0596]	-0.076* [0.0410]
Fixed Effects	Y	Y	Y	Y
Adj. R-squared	0.226	0.254	0.257	0.228
Firms	918	956	956	956
Observations	49,723	52,888	52,888	52,888

Table 3.7: Disagreement and Salience of information signals

Notes: The table reports the Disagreement regression results and the moderating effect of the salience of information signals of StockTwits. Salience is divided into two groups, i.e., Reach, which represents the magnitude of access to ideas posted by investors on StockTwits; Social Media Attention (SMA) represents alternative attention proxies of StockTwits. These salience groups are further divided into different subgroups, which are defined in appendix 3.2. The dependent variable is Return Synchronicity, which is calculated using the Carhart (1997) four-factor model. Disagreement is derived from the investors' recommendations on StockTwits. All right-hand side variables are standardized. The sample consists of 956 firms with 52,888 firm-month observations for the sample period between 2013 - 2017. The regressions are estimated using time, firm, and industry fixed effects. Variable definitions are presented in appendix 3.2. *, **, and *** represent significance at 10%, 5%, and 1%, respectively. The standard errors are reported in brackets and are robust to heteroscedasticity and clustered at the firm and month levels.

	1	**SMA**		
		2	**3**	**4**
	Network	**Ideas**	**Popularity**	**Discussion**
Disagreement	-0.067*** [0.0120]	-0.059*** [0.0120]	-0.066*** [0.0102]	-0.046*** [0.0099]
Salience	-0.157*** [0.0174]	-0.208*** [0.0235]	-0.120*** [0.0228]	-0.132*** [0.0171]
*Disagreement * Salience*	-0.078*** [0.0084]	-0.075*** [0.0093]	-0.054*** [0.0146]	-0.026*** [0.0084]
Media Coverage	-0.174*** [0.0225]	-0.148*** [0.0226]	-0.199*** [0.0227]	-0.196*** [0.0237]
Analyst Coverage	0.120*** [0.0150]	0.112*** [0.0148]	0.115*** [0.0147]	0.115*** [0.0144]
Leverage	-0.087*** [0.0236]	-0.034 [0.0233]	-0.050** [0.0213]	-0.066** [0.0242]
Adv/Sales	-0.013* [0.0071]	-0.013* [0.0072]	-0.015* [0.0075]	-0.015* [0.0072]
Market/Book Ratio	-0.031** [0.0112]	-0.040*** [0.0108]	-0.041*** [0.0109]	-0.036*** [0.0112]
Firm Size	0.195*** [0.0590]	0.182*** [0.0563]	0.197*** [0.0574]	0.195*** [0.0561]
ROA	0.053*** [0.0156]	0.054*** [0.0149]	0.056*** [0.0153]	0.054*** [0.0154]
Earnings Volatility	0.016** [0.0052]	0.014** [0.0049]	0.014** [0.0052]	0.015** [0.0053]
Sales Growth	0.056* [0.0303]	0.074** [0.0325]	0.069* [0.0314]	0.059* [0.0317]
Real GDP_{t-1}	-0.071 [0.0407]	-0.069 [0.0397]	-0.078* [0.0406]	-0.073* [0.0402]
Fixed Effects	Y	Y	Y	Y
Adj. R-squared	0.235	0.247	0.237	0.236
Firms	956	956	956	956
Observations	52,888	52,888	52,888	52,888

Table 3.8: Disagreement and Heterogeneity of Investors

Notes: The table reports the regression results of the Disagreement and the moderating effect of the heterogeneity of investors on StockTwits. The dependent variable is Return Synchronicity, which is calculated using the Carhart (1997) four-factor model. Disagreement is derived from the investors' recommendations on StockTwits. The heterogeneity of investors is further divided based on the investors' experience and investment approaches. Model (1) presents the regression results based on overall investors, where investors are defined as the number unique of investors posting ideas on StockTwits while discussing the sample firms. Model (2) and (3) present the regression results based on a within-group disagreement between investors with self-disclosed investment experience and investment approaches, respectively. The difference test is the p-value associated with the t-test for differences in the coefficients of Disagreement between Professional and Novice in the model (2) and Momentum and Value investors in the Model (3), respectively. The sample consists of 956 firms with overall 52,888 firm-month observations for the sample period between 2013 - 2017. The regressions are estimated using time, firm, and industry fixed effects. Variable definitions are presented in appendix 3.2. *, **, and *** represent significance at 10%, 5%, and 1%, respectively. The standard errors are reported in brackets and are robust to heteroscedasticity and clustered at the firm and month levels.

	1	**2**	**3**
Disagreement$_{Overall}$	-0.061***		
	[0.0120]		
Investors	-0.212***		
	[0.0261]		
*Disagreement$_{Overall}$ * Investors*	-0.079***		
	[0.0096]		
Disagreement$_{Professional}$		-0.054***	
		[0.0070]	
Disagreement$_{Intermediate}$		-0.048***	
		[0.0087]	
Disagreement$_{Novice}$		-0.037***	
		[0.0073]	
Disagreement$_{Momentum}$			-0.062***
			[0.0075]
Disagreement$_{Technical}$			-0.051***
			[0.0092]
Disagreement$_{Fundamental}$			-0.044***
			[0.0056]
Disagreement$_{Value}$			-0.027***
			[0.0051]
Media Coverage	-0.152***	-0.207***	-0.193***
	[0.0225]	[0.0232]	[0.0239]
Analyst Coverage	0.113***	0.119***	0.120***
	[0.0149]	[0.0148]	[0.0148]
Leverage	-0.031	-0.086***	-0.085***
	[0.0232]	[0.0238]	[0.0244]
Adv/Sales	-0.013*	-0.013*	-0.013*
	[0.0071]	[0.0068]	[0.0070]
Market/Book Ratio	-0.040***	-0.035***	-0.033**
	[0.0107]	[0.0112]	[0.0114]
Firm Size	0.184***	0.198***	0.194***
	[0.0566]	[0.0579]	[0.0585]
ROA	0.054***	0.053***	0.054***
	[0.0150]	[0.0160]	[0.0159]
Earnings Volatility	0.014**	0.015**	0.015**
	[0.0050]	[0.0058]	[0.0060]
Sales Growth	0.077**	0.052	0.053
	[0.0327]	[0.0303]	[0.0305]
Real GDP$_{t-1}$	-0.070	-0.076*	-0.076*
	[0.0399]	[0.0408]	[0.0406]
Diff (p-value)		(0.1347)	(0.003)***
Fixed Effects	Y	Y	Y
Adj. R-squared	0.247	0.232	0.235
Firms	956	956	956
Observations	52,888	52,888	52,888

Table 3.9: Disagreement and alternative proxies of Return Synchronicity

Notes: The table reports the regression results of the Disagreement and alternative proxies of Return Synchronicity. In Model (1), Return Synchronicity is calculated using the Carhart (1997) four-factor model and using the Adjusted R-squared. In Model (2) and (3), we use Morck et al. (2000) approach to calculate Return Synchronicity based on R-squared and adjusted R-squared. In Model (4), we follow Peng and Xion's (2006) approach to calculate Return Synchronicity as a time series of Pearson correlation coefficients. All right-hand side variables are standardized. The sample consists of 956 firms, whereas firm-month observations vary based on the difference of approaches. The sample period is 2013 – 2017. The regressions are estimated using time, firm, and industry fixed effects. Variable definitions are presented in appendix 3.2. *, **, and *** represent significance at 10%, 5%, and 1%, respectively. The standard errors are reported in brackets and are robust to heteroscedasticity and clustered at the firm and month levels.

	1	**2**	**3**	**4**
	FF4-Adj	**MYY**	**MYY-Adj**	**CORR**
Disagreement	-0.059*** [0.0122]	-0.027** [0.0095]	-0.026* [0.0119]	-0.024*** [0.0030]
Media Coverage	-0.241*** [0.0291]	-0.170*** [0.0199]	-0.187*** [0.0223]	-0.054*** [0.0054]
Analyst Coverage	0.096*** [0.0149]	0.089*** [0.0138]	0.078*** [0.0158]	0.031*** [0.0041]
Leverage	-0.127*** [0.0283]	-0.074** [0.0276]	-0.088** [0.0346]	-0.029*** [0.0060]
Adv/Sales	-0.019*** [0.0060]	-0.002 [0.0058]	-0.007 [0.0046]	-0.004*** [0.0008]
Market/Book Ratio	-0.028 [0.0162]	-0.030** [0.0115]	-0.039** [0.0132]	-0.005 [0.0042]
Firm Size	0.217*** [0.0689]	0.165** [0.0544]	0.167** [0.0662]	0.056*** [0.0131]
ROA	0.048** [0.0216]	0.045** [0.0199]	0.056* [0.0257]	0.010** [0.0040]
Earnings Volatility	0.015 [0.0114]	0.020* [0.0101]	0.027*** [0.0085]	0.002 [0.0016]
Sales Growth	0.051 [0.0345]	0.069** [0.0271]	0.087** [0.0358]	0.016* [0.0077]
Real GDP_{t-1}	-0.083 [0.0503]	-0.053 [0.0369]	-0.058 [0.0431]	-0.023* [0.0117]
Fixed Effects	Y	Y	Y	Y
Adj. R-squared	0.168	0.342	0.286	0.270
Firms	956	956	956	956
Observations	44,548	52,686	47,266	52,916

Table 3.10: Return Synchronicity and Disagreement based on alternative prediction models

Notes: The table reports the regression results of the Disagreement and Return Synchronicity. The dependent variable is Return Synchronicity, which is calculated using the Carhart (1997) four-factor model using R-squared, adjusted R-squared, and time series of Pearson correlation coefficient, respectively. Disagreement is derived from the users' recommendations predicted using Support Vector Machine and Maximum Entropy models. All right-hand side variables are standardized. The sample consists of 956 firms with variable firm-month observations in the Model (1) – (6) for the sample period between 2013 - 2017. The regressions are estimated using time, firm, and industry fixed effects. Variable definitions are presented in appendix 3.2. *, **, and *** represent significance at 10%, 5%, and 1%, respectively. The standard errors are reported in brackets and are robust to heteroscedasticity and clustered at the firm and month levels.

	Support Vector Machine			**Maximum Entropy**		
	1	**2**	**3**	**4**	**5**	**6**
	FF4	**FF4-Adj**	**CORR**	**FF4**	**FF4-Adj**	**CORR**
Disagreement	-0.043**	-0.050***	-0.022***	-0.042***	-0.049***	-0.022***
	[0.015]	[0.015]	[0.004]	[0.013]	[0.013]	[0.003]
Media Coverage	-0.228***	-0.240***	-0.053***	-0.230***	-0.242***	-0.054***
	[0.023]	[0.028]	[0.005]	[0.024]	[0.029]	[0.005]
Analyst Coverage	0.116***	0.096***	0.031***	0.116***	0.095***	0.031***
	[0.015]	[0.015]	[0.004]	[0.014]	[0.015]	[0.004]
Leverage	-0.095***	-0.126***	-0.028***	-0.095***	-0.125***	-0.028***
	[0.024]	[0.028]	[0.006]	[0.024]	[0.028]	[0.006]
Adv/Sales	-0.015*	-0.019**	-0.004***	-0.015*	-0.019**	-0.004***
	[0.007]	[0.006]	[0.001]	[0.007]	[0.006]	[0.001]
Market/Book Ratio	-0.032**	-0.029*	-0.006	-0.032**	-0.028	-0.005
	[0.011]	[0.016]	[0.004]	[0.011]	[0.016]	[0.004]
Firm Size	0.200***	0.213**	0.054***	0.202***	0.216***	0.055***
	[0.058]	[0.069]	[0.013]	[0.058]	[0.069]	[0.013]
ROA	0.054***	0.048**	0.010**	0.054***	0.049**	0.010**
	[0.016]	[0.022]	[0.004]	[0.016]	[0.022]	[0.004]
Earnings Volatility	0.016**	0.015	0.002	0.016**	0.015	0.003
	[0.006]	[0.011]	[0.002]	[0.006]	[0.011]	[0.002]
Sales Growth	0.051	0.051	0.016*	0.050	0.049	0.015*
	[0.030]	[0.034]	[0.008]	[0.030]	[0.035]	[0.008]
Real GDP_{t-1}	-0.076*	-0.083	-0.023*	-0.076*	-0.084	-0.023*
	[0.041]	[0.050]	[0.012]	[0.041]	[0.050]	[0.012]
Fixed Effects	Y	Y	Y	Y	Y	Y
Adj. R-squared	0.227	0.168	0.269	0.227	0.168	0.269
Firms	956	956	956	956	956	956
Observations	52,888	44,548	52,916	52,888	44,548	52,916

Appendix 3.1

StockTwits Recommendation Predictions and Text Analysis

For recommendation predictions of StockTwits ideas, we use the Random Forest Decision Trees model (RFDT). For RFDT to work as an ensemble, decision trees are created based on the impurity criterion. In our case, we use Entropy as the impurity criterion.

$$Entropy = -\sum_{i=1}^{L} f_i \log(f_i)$$

(1)

In equation (1), f_i is the frequency of label *i* at a node and *L* is the total number of unique labels. After calculating Entropy, the next step is to measure information gain, which is the metric that measures the expected reduction in the level of impurity in a given data set.

$$IG_{j,k} = Entropy_j - Entropy_{j,k}$$

(2)

In equation (2), $IG_{j,k}$ is the information gain from the given sample, j is the target value, k is split features, $Entropy_j$ is Entropy calculated for the target value and $Entropy_{j,k}$ is Entropy after the data are split based on *k* features. However, it is pertinent to note that even if the impurity criterion is changed to Gini impurity, our results remain constant, and there is no change in prediction outcome scores, such as AUC measures or F1 scores. To start applying the RFDT model, for each decision tree, the importance of the node is calculated based on the impurity criterion, as follows:

$$\theta_j = W_j C_j - W_j^{left} C_j^{left} - W_j^{Right} C_j^{Right}$$

(3)

Where θ_j is the importance of node *j*, W_j is the weighted number of samples reaching node *j*, C_j is the impurity value of node *j*. Similarly, $W_j^{left} C_j^{left}$ presents the values from the left node and $W_j^{Right} C_j^{Right}$ presents the values from the right node. In the next step, the importance of each feature *i* on the decision tree is calculated as follows:

$$\pi_i = \frac{\sum_{j:node\ j\ split\ on\ feature\ i} \theta_j}{\sum_{a\ \epsilon\ all\ nodes} \theta_a}$$

(4)

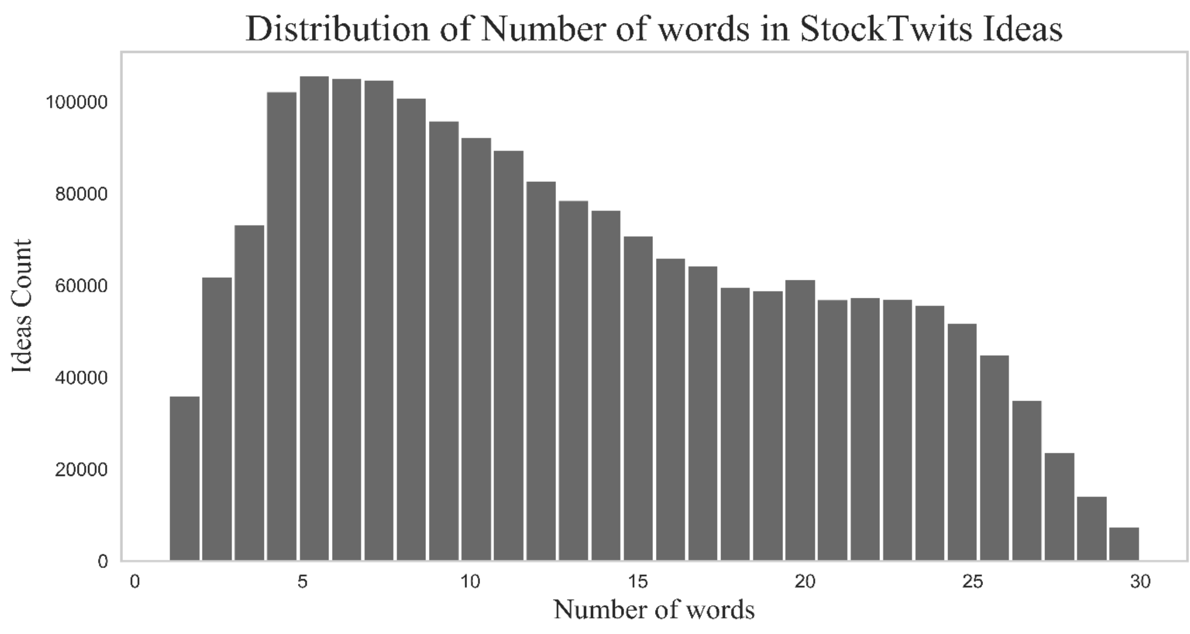
Where π_i is the importance of feature i, θ_j is the importance of node j from equation (3), and θ_a is the importance of all the nodes in a given tree. These values are then normalised, ranging between 0 and 1. Finally, overall feature importance i in all trees γ at the random forest level is calculated as an average of the features of all the decision trees.

$$\lambda_i = \frac{\sum_{\gamma \in all\ trees} \pi_i}{\sum DT} \tag{5}$$

In equation (5), λ_i is the importance of feature i calculated from all trees that are part of the random forest model, π_i is the normalized value of feature importance calculated in equation (4), and DT is the total number of decision trees that are part of the random forest ensembling process. Figure F2 in the internet appendix presents a sample decision tree.

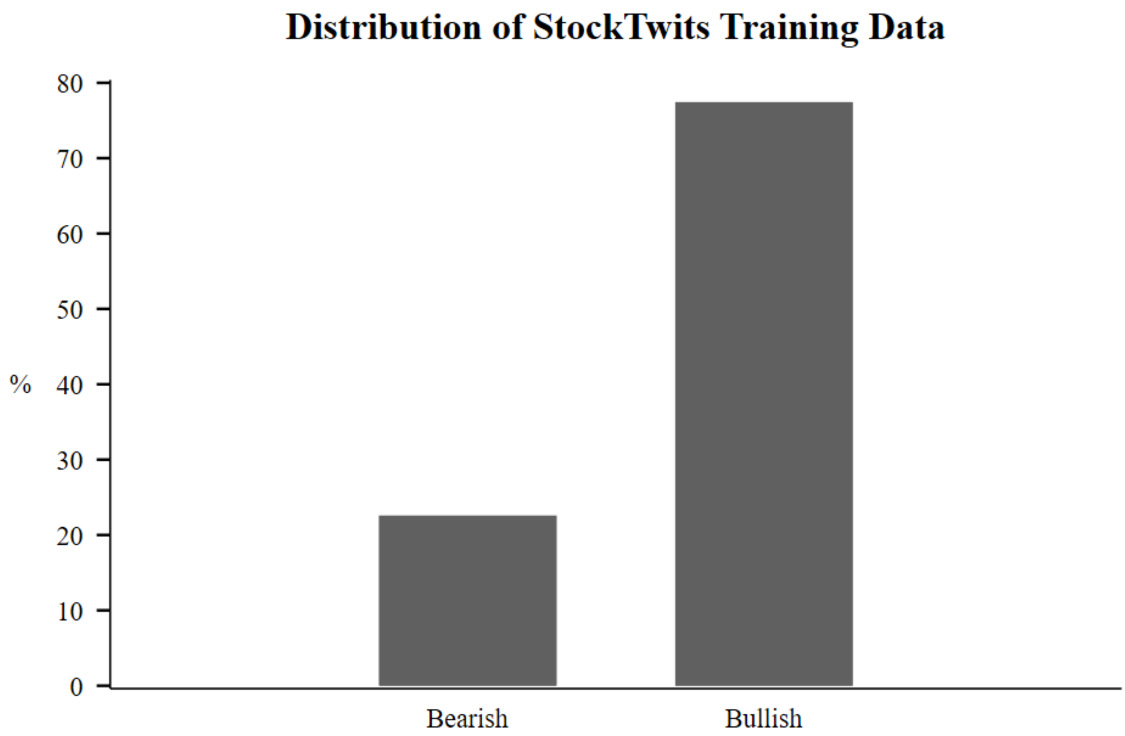
Text Analysis and Tokenisation

We use Baziotis et al. (2017) *Ekphrasis* library for text analysis. *Ekphrasis* is trained on Wikipedia and more than 330 million tweets. *Ekphrasis* also translates regular expressions used in StockTwits ideas in the form of emojis by using its social tokenizer. Table A3 in the internet appendix presents randomly selected ideas along with the bag of words. To further explore the quality of our StockTwits dataset, we plot the distribution of the number of words in StockTwits ideas. Overall, the average number of words in the distribution is 13.01, with a median of 12 words. The standard deviation of the distribution is 7.55 words, whereas the minimum length of ideas is 1 and the maximum length of ideas is 115 words.



Training Prediction Models

We train our prediction model based on self-labelled[85] ideas from investors on StockTwits. The size of the training data set plays a vital role in the accurate prediction of investors' recommendations. Unlike Antweiler and Frank (2004), who used 1,000 messages for training, Giannini et al. (2019), who used 2,000 Twitter posts for training, and Cookson and Niessner (2019), who used 472,368 StockTwits ideas, the training data set used in our prediction models is 1.92 million StockTwits ideas. The distribution of our training data set is presented in the following figure.



Prediction Accuracy

In the RFDT model, our F1 score is 89%, and the AUC[86] score is 79%. Considering the importance of k-fold cross-validation, we test the validity of our model's prediction using 10-fold cross-validation (CV). Our cross-validation accuracy is 82% and falls well within the range of robust prediction models. We select the best model based on the F1 score and CV accuracy. In addition to the RFDT model, we also use Support Vector Machine (SVM) and

[85] This is important to control subjective bias when a data set is hand-labelled and then calibrated in prediction models.
[86] Area Under the Curve

Maximum Entropy (MaxEnt) to predict investors' recommendations. We use these models to check the robustness of our results in section 6 of this study. The table below presents the results from three prediction models we use in this study.

Recommendation Prediction Models

Prediction Models	AUC	Accuracy	F1 Score
Random Forest Decision Trees	0.79	0.82	0.89
Support Vector Machine	0.81	0.80	0.87
Maximum Entropy	0.77	0.74	0.82

Following table presents the sample list of bullish and bearish words in the investors' recommendations on StockTwits after the sentiment analysis based on the RFDT model.

Bullish and Bearish words after the Sentiment Analysis

Bullish		Bearish	
Aggressive	Thestreet	Announcement	Shoulders
Beat	Trade	Attention	Sinking
Earnings	Trades	Away	Spike
Forecast	Trailing	Buyback	Split
Fundamentals	Trending	Confidence	Spot
Gains	Tripled	Consolidation	Spread
Historical	Undervalued	Credit	Strategy
Hold	Upgrade	Decrease	Technology
Lawsuits	Uptrend	Disappointed	Topped
Leadership	Value	Downgraded	Trouble
Legal	Volatility	Downward	Tumble
Lock	Voting	Freefall	Understand
Long	Warren	Future	Unfortunately
Patent	Watching	Information	Unload
Psychological	Wave	Loss	Upcoming
Published	Wednesday	News	Vertical
Ratings	Weekend	Outlook	Volatility
Ratio	Welldone	Pullback	Wait
Research	Winner	Put	Wiped
Revisions	Wires	Resistance	Wish
Signal	Witnessing	Risky	Worry
Special	Worth	Rumors	Worst
Stochastic	Yield	Run	Worthless
Surprise		Sell	Wrong
Technology		Short	

Appendix 3.2: Variables Definitions

Variables	Data Source	Frequency	Description
			StockTwits Variables
Disagreement	StockTwits	Monthly	Disagreement is calculated by predicting the investors' recommendations using the Random Forest Decision Trees analysis and then following Cookson and Niessner (2019). The disagreement level ranges between 0 as agreement and 1 as a complete disagreement between the investors on StockTwits.
Ideas			Natural logarithm of the number of ideas posted by the investors on StockTwits while discussing the sample firms.
Investors			Natural logarithm of the number of distinct users posting on StockTwits while discussing the sample firm.
Proximity			Natural logarithm of the number of distinct users from the same US State where the firm headquarters is located, posting on StockTwits while discussing the sample firm.
Network			The sum of distinct investors' followers who post ideas while discussing the sample firms.
Popularity			Natural logarithm of the number of likes a user post receives while discussing the sample firm.
Discussion			Natural logarithm of the number of replies a post receives while discussing the sample firm.
$Revisions_t$			Natural logarithm of the sum of the number of times a distinct investor revises her recommendations (Bullish-to-Bearish) or vice versa daily and then aggregated at monthly frequency.
$Revisions_{t-1}$			Natural logarithm of the lagged value of the sum of the number of times a distinct investor revise her recommendations (Bullish-to-Bearish) or vice versa daily and then aggregated at a monthly frequency
			Media Coverage
Breaking News	Ravenpack	Monthly	Natural logarithm of number of breaking news covered in the media related to the firm.
Full Articles			Natural logarithm of the number of detailed articles and editorials published in the media while discussing the firm.
Press Release			Natural logarithm of the number of press releases issued by the firm.

Variables	Data Source	Frequency	Description
		Firm-Level Characteristics	
Return Synchronicity	CRSP	Monthly	Return synchronicity is calculated based on the value of the coefficient of determination (R^2). To derive the value of the coefficient of determination, we use Carhart's (1997) four-factor model and take the log transformation of the coefficient of determination.
Firm Size		Quarterly	Natural logarithm of Market Capitalisation of the Firm.
Analyst Coverage	I/B/E/S	Monthly	Natural logarithm of Number of Analysts covering the firm.
Leverage		Quarterly	Total long-term debt scaled by total assets.
Adv/Sales		Quarterly	Advertising to sales ratio of the sample firms.
Market/Book Ratio		Quarterly	The market value of equity is scaled by the book value of the total assets.
Sales Growth		Quarterly	Natural logarithm of firm sales.
ROA		Quarterly	Income before extraordinary items scaled by total assets.
Earnings Volatility	Compustat	Quarterly	Standard Deviation of the ratio between income before extraordinary items in the current quarter and total assets in the previous quarter.
Firm Opacity		Yearly	Discretionary accruals are used as a proxy of firm opacity or measure of firms' accrual quality calculated based on the extended Jones (1991) model.
Diversity		Yearly	The natural logarithm of the number of business segments of the firm is multiplied by the number of geographic segments.
Competition		Yearly	Industry competition is based on firm-level assets calculated using the Herfindahl-Hirschman Index.
Insider Trading	Thomson Reuters	Monthly	Insider trading is calculated as the difference between the buy and sell insider trades scaled by the total insider trades in a given month of a sample firm.
Real GDP	US Bureau of Economic Analysis	Monthly	Real Gross Domestic Product (GDP) of the United States.

Appendix 3.3: Selected Ideas from StockTwits

afernandez321 Bullish 3/13/15, 10:43 AM

$TSLA many say we are in trouble. I completely disagree. This 61.8 Fib indicator is holding up very well.

LDrogen 6/13/12, 03:01 PM

obviously the market says $CTCT paid too much today, at 30X revs I wouldn't disagree

OphirGottlieb Bullish 8/9/16, 10:25 PM

Lifetime

$SBUX Analysts getting bearish. I disagree

ibdt 5/24/12, 03:09 PM

@BlackStockMarker I disagree that $AAPL strong. Looking at their history they are weak // Really? From June 2007 $AAPL is _UP_ 400%. Weak?!?

amigobulls Bearish 4/7/14, 03:59 PM

$FB Worth a deeper look for the GARP investor Agree or disagree? amigobulls.com/stocks/FB/an...

swansong Bearish 7/25/14, 07:32 PM

$GTAT Had it closed above $15.25+ today, I would disagree with @Coinster. Have to agree though, EOD was weak today, and for the time being

lamonicabuzz 8/4/15, 03:01 PM

$CFR's Evans: "We live in a time where Washington thinks it knows more than free enterprise. I disagree strongly with that."

jfahmy 4/12/13, 12:13 PM

As much as I respect Gundlach, I disagree with him on his $CMG short. I think it's a great long-term growth story (no position currently).

Stocktwits 10/14/15, 12:49 PM

The analyst who downgraded $GPRO. Not saying agree/disagree. But... "@Trader4fun: $GPRO Erinn Murphy, certainly a "top" analyst..."

MountainPhantom Bullish 9/14/16, 03:13 PM

@BiotechTrades The three 20K share blocks that were bought at 11.55 disagree with you - accumulation underway! $DVAX

The_Analyst Bullish 4/11/16, 10:41 AM

Sympathy for $OSTK CEO Byrne. Announces medical leave of absence and stock pops. Ouch. (not that I disagree w/market reaction)

FinancialJuice 7/18/13, 04:51 PM

Pepsico $PEP investor BlackRock says disagrees with Nelson Peltz on proposed purchase by Co. of Mondelez

JFinDallas Bearish 8/31/15, 11:33 AM

Don't know about the $10 PT but can't say I disagree with Citron on $W... citronresearch.com/wp-conte...

bradloncar 5/29/13, 08:28 AM

I love it when analysts disagree. Pacific Crest lowers $ENOC to sector perform, Raymond James upgrades it to strong buy.

Appendix 3.4: Selected newspaper articles discussing the Labour Disputes

Reuters News
21 March 2015

U.S. mediators report tentative deal in Honeywell Illinois dispute
(Reuters) - A tentative agreement has been reached in a labor dispute at a Honeywell International Inc uranium conversion plant in Illinois, potentially ending a work stoppage that began last August, federal mediators and the company said on Saturday.
Production and maintenance employees at the Metropolis, Illinois, plant who are members of United Steelworkers Local 7-669 have been locked out since Aug. 2 after a three-year contract expired.
"The company is pleased to have reached a tentative agreement on a new, three-year contract. The union has indicated that its membership will vote on the proposed contract on Wednesday or Thursday," company spokesman Peter Dalpe said.
In a statement, the U.S. Federal Mediation and Conciliation Service (FMCS) said it asked the parties to return to the bargaining table with federal mediation on Wednesday. "Working under the auspices of FMCS mediators, the union and employer reached a tentative agreement after many hours of intense talks," the statement said.
The mediation service said the tentative agreement is subject to ratification. It provided no details on the agreement.
"At the request of the federal mediator who assisted in the bargaining this week, the company is not commenting on the specific terms of the proposed contract at this time, except to say that we are hopeful the agreement will be ratified," Dalpe said.
Honeywell's Metropolis plant is the only U.S. facility that converts uranium oxide into uranium hexafluoride, which is then enriched to be used as fuel in nuclear power plants.
"We are extremely pleased that we were able to help the union and employer overcome their differences in these challenging and complex negotiations," Allison Beck, FMCS acting director, said in the statement.
"Representatives from both sides were willing to put in the long hours that were required, and ultimately they were successful in reaching a mutually acceptable agreement," Beck said.

The Wall Street Journal Online

5 September 2014
Silicon Valley Companies Appeal Wage Decision; Attorneys for Apple, Google, Intel and Adobe Systems Say Judge Committed Legal Error
Four big Silicon Valley companies appealed a judge's decision to throw out a proposed settlement with 64,000 current and former workers who claim the companies collaborated to suppress wages.
Attorneys for Apple Inc., Google Inc., Intel Corp. and Adobe Systems Inc. said U.S. District Judge Lucy Koh "committed clear legal error" in rejecting a proposed $324.5 million settlement that the companies had reached with attorneys for the workers.
They appealed the ruling to the Ninth U.S. Circuit Court of Appeals, based in San Francisco, asking the appeals court to order Judge Koh to approve the settlement.
Judge Koh rejected the proposed settlement on Aug. 8, saying it was too little to compensate the workers, and took the unusual step of suggesting $380 million would be more in line with the previous settlements with three other companies in the case.
The workers claim that the companies agreed not to recruit each other's workers, suppressing wages from 2005 to 2009. The case is based on a 2011 Justice Department case that also involved Intuit Inc., and the Pixar and Lucasfilm units of Walt Disney Co., which settled previously with the workers.
In their appeal, the four remaining companies said the judge "applied a mechanical formula that overrode sensitive judgments of the class's [workers] own counsel." They said the ruling "will inflict significant harm on all parties and the class action procedure."
Attorneys for the workers didn't respond to requests for comment. Intel and Google declined to comment. Apple, Adobe, and Judge Koh didn't immediately respond to requests for comment.
Judge Koh on Thursday said she plans to start a trial in the case on Jan. 12, and ordered a pretrial hearing on Dec. 18. On Wednesday, attorneys filed a document saying they had resumed mediation with former U.S. District Judge Layn Phillips.
A trial could put top Silicon Valley executives, such as former Google Chief Executive Eric Schmidt and company founders Sergey Brin and Larry Page, on the witness stand. The companies' lawyers have told the court they might call as many as 70 witnesses, the workers' attorneys said they might call 37. Neither side has apparently made its list public.
The case also revealed emails from the late Steve Jobs that made headlines because of the former Apple CEO's personal involvement in opposing other companies' attempts to hire Apple employees.

Chapter 4

4. To revise is to devise? Sentiment Oscillations on Social Media and Crash Risk

When the facts change, I change my mind . . .

- John Maynard Keynes

4.1. Introduction

Social media has recently emerged as a powerful information-sharing platform by empowering the wisdom of crowds to influence financial markets. Recently, the social media-driven trading frenzy for GameStop (GME) has attracted regulators and professional money managers to acknowledge the significance of such platforms. Previous literature has mainly focused on investigating the role of social media to predict volumes, returns, investors' trading behaviors in financial markets, and as information dissemination channels to reduce information asymmetry.[87] However, there is a lack of evidence to investigate the influential role of social media to predict extreme movements of stock prices in financial markets. Our study fills this gap by examining the role of investor-oriented social media platforms to facilitate investors to predict firm-specific crash risk.

Stock price crashes are rare to market events and are of particular interest to investors and regulators in financial markets. Previous studies (Hutton et al., 2009; Jin & Myers, 2006; Kothari et al., 2009) present evidence that managers endeavor to withhold negative firm-specific information due to their career concerns and a range of other benefits, resulting in stock price crashes. However, such information cannot be withheld for an extended period and eventually reaches a tipping point. Therefore, a sudden release of bad news causes stock prices to crash. Recent studies provide further evidence that the higher sensitivity of CFOs' wealth to firm stock price (Kim et al., 2011b), increased short-selling (Callen & Fang, 2015b), financial reporting conservatism (Kim & Zhang, 2016), increased stock price liquidity (Chang et al., 2017), increased CEO power (Al Mamun et al., 2020) and an increase in firm-level distress risk (Andreou et al., 2021), among others, are positively associated with crash risk. The existing

[87] For example, Bollen et al. (2011) highlight the significance of investors' sentiment on Twitter to predict returns; Chen et al. (2014) using data from SeekingAlpha, a popular social media platform, present how investors' opinions can predict volumes and returns; Cookson and Niessner (2019) provide evidence that disagreements among investors on StockTwits can predict abnormal volumes; Blankespoor et al. (2014) highlight the significance of a social media platform (Twitter) to disseminate corporate press releases and reduce information asymmetry. For a detailed review of the role of social media to facilitate corporate communication, see Blankespoor (2018).

literature provides unified evidence that managers' bad news hoarding increases stock price crash risk; and eventually, crashes occur when such information leaks out.

Investor-oriented social media platforms have transformed the way investors interact with each other. For example, without having a physical presence while sifting through millions of threads, investors can interact with each other using such platforms. Therefore, these platforms assist investors in many ways. First, by providing a diversity of opinions since there are no geographic or demographic restrictions. Second, encouraging investors to learn investment tricks and techniques by following influential investors in a different or similar area of interest. Third, providing firm-specific information from, and value-added analysis by, other investors, thus keeping the cost of accessing such information very low. Last, providing opportunities for investors to become influencers on such platforms by sharing firm-specific analyses based on one's knowledge and investment skills. Motivated by these characteristics of investor-oriented social media platforms, we investigate the role that sentiment oscillations play on one of the most popular social media platforms for investors, i.e., StockTwits,[88] to predict firm-level future crash risk. We define sentiment oscillations as revisions in investors' sentiment while discussing specific sample firms on StockTwits.

Investor sentiment plays a pivotal role in understanding the various behavioral and psychological factors that drive financial markets (Barberis, 2018). Rubinstein (1993) argues that market participants have distinct preferences and the ability to consume information. Therefore, the interpretation of information signals by such market participants differs from one participant to another.[89] To understand investors' differential interpretations, Harris and Raviv (1993) and Kandel and Pearson (1995) present differential-interpretation hypotheses, suggesting that investors interpret information signals based on their preferences and economic models; consequently, such differential interpretations can predict volumes and returns.

In the behavioral finance literature, researchers like Shiller (1980) provide evidence that rational models in finance cannot forecast a firm's future cash flows. Therefore, there is a clear need to develop new behavioral models to explain the fluctuations in financial markets. Similarly, Shleifer and Summers (1990) provide an alternative explanation of the efficient market hypothesis in the presence of rational and less rational investors and suggest the need

[88] StockTwits is by far the largest social media platform for investors. As of March 2021, it has more than 5 million registered users, more than 7 million monthly messages per month and 5 million monthly visitors.

[89] We assume that investors' sentiment is a by-product of the interpretation of information signals by investors. This is in line with previous studies such as Barberis and Thaler (2003), Antoniou et al. (2013); Da et al. (2014); Hirshleifer (2015) and Barberis (2018).

to investigate the role of investors' sentiment and limits to arbitrage. Although these studies highlight the role that investor sentiment plays in financial markets, they do not focus on providing behavioral explanations of market crashes. Hong and Stein (2003) present a theory of market crashes based on differences of opinion in financial markets and argue that information asymmetry in financial markets can explain market crashes.

Our hypothesis is motivated by Hong and Stein (2003) *difference-of-opinion* theory, which provides a behavioral explanation of crash risk. Following Hong-Stein's two-period model, we assume a continuum of heterogeneous investors in a large social network that offers investors unique opportunities to update their sentiment based on available information. At time 1, both optimistic and pessimistic investors analyze available information signals and, based on their interpretation, decide to take a position (*optimists*) or stay out of the market (*pessimists*). At time 2, when updated information arrives in the market, based on their analysis and interpretation of information signals, investors can stay put if updated information signals support their economic models or revise their sentiment. Hong-Stein suggests that the arrival of bad news convinces optimistic investors to revise their sentiment (*bullish-to-bearish*).

Similarly, pessimists do not revise their sentiment on the arrival of bad news since they are already out of the market. Under such circumstances, revisions in investors' sentiment should predict the negative skewness of returns as more information dribbles out, which coincides with pessimist investors' economic models. The emergence of social media platforms for investors provides a unique laboratory to investigate the role of investor sentiment and test the application of *difference-of-opinion* theory.

We employ three continuous and three dichotomous crash risk measures to examine the association between sentiment oscillations and crash risk. Such a variety of crash risk measures allows our analysis to capture the magnitude of the negative skewness of returns using continuous measures and the predictability of extreme movements based on dichotomous measures. Investors on StockTwits voluntarily disclose their sentiments (*Bullish/Bearish*). To ensure that we have investors' sentiment for all the ideas posted on StockTwits while discussing the sample firms, we predict investors' sentiment using the Maximum Entropy (MaxEnt) model. First, we train the data set using more than one million self-labeled ideas from StockTwits, and then use those training data to predict investors' sentiment on StockTwits. Our sentiment measures are robust to 10-fold cross-validation and an automatic feature selection model. Our final data set contains more than 13 million StockTwits ideas posted by 166,949 unique investors worldwide. We define sentiment oscillations as changes in investor sentiment from *bullish-to-bearish* and vice versa and assign weights to sentiment oscillations based on the number of individual investors' followers. This gives us a unique opportunity to account for the "*Contagion effect*," as explained by Hong and Stein (2003).

Our results show that there is a positive association between sentiment oscillations and future crash risk. Using a continuous dependent variable ($NCSKEW_t$), the results indicate that one standard deviation increase in sentiment oscillations results in 0.04 standard deviation increase in crash risk, which equates to 53.33% of the unconditional mean value of negative conditional skewness of returns ($NCSKEW_t$). Similarly, based on a dichotomous dependent variable (*CRASH309*), the results show that one standard deviation increase in sentiment oscillations results in a 0.28% increase in the probability of future crash risk. These results suggest that the association between sentiment oscillations is statistically and economically significant and are consistent with our hypothesis that firms with higher sentiment oscillations are associated with higher crash risk in the future.

These findings highlight the transactional role of investor-oriented social media platforms by suggesting that an increase in sentiment oscillations indicates that investors update their economic models based on available information and eventually diffuse their information signals to a large social network of investors, thus increasing the magnitude of the impact of sentiment oscillations to predict future crash risk. These results remain robust after using a large set of control variables (consistent with prior literature) and using firm, industry, and time fixed effects. Our results are consistent with prior studies predicting firm-level future crash risk. However, to the best of our knowledge, this is the first study to investigate the role of investor-oriented social media platforms to assist investors to predict future crash risk.

There are two key mechanisms through which we need to test the validity of our hypothesis. First, Ak et al. (2016) present evidence that more than 70% of crashes occur due to earnings surprises.[90] Similarly, Skinner and Sloan (2002) suggest that earnings announcements are one of the common causes of stock price crash risk. Motivated by these studies, we extend our analysis to test the validity of our results when there are quarterly earnings announcements. We isolate our sample based on firm-specific stock price crash and earnings announcements occurring in the same month. Therefore, dividing our firm-month sample into earnings

[90] The rest of the causes include preannouncement surprises and the outcomes of clinical trials in the case of the healthcare sector, among others.

announcement (QEA) and non-earnings announcement (Non-QEA) months and estimate the regression following our baseline model. The results suggest that the predictive power of sentiment oscillation is double for the *QEA* sample in contrast to the *Non-QEA* sample. These results are in line with prior literature and especially An et al. (2020), who suggest that using earnings announcement is a critical test to mitigate the impact of unexpected market events on crash risk.

Second, Kandel and Pearson (1995) suggest that the heterogeneity of investors plays a pivotal role in financial markets, i.e., an increase in heterogeneity suggests diversity among investors in financial markets. Hong and Stein (1999) and Hong and Stein (2007) highlight the role of market segmentation and specializations by suggesting that some investors in financial markets act as frontrunners to access value-relevant information. Motivated by these studies, we investigate how investors' heterogeneity accentuates the power of sentiment oscillations to predict firm-level future crash risk. Investors on StockTwits voluntarily disclose their investment experience. We divide the investors into two groups,[91] i.e., *Professional* and *Novice,* and measure each group's sentiment oscillations[92] and estimate regressions using our baseline model. Our results show that the stand-alone sentiment oscillations of both professional and novice investors can predict future crash risk. However, when combined, the overall comparison suggests that professional investors subsume the impact of novice investors' sentiment oscillations to predict future crash risk. These results are consistent with the previous literature and further support the validity of our results, thus suggesting that the heterogeneity of investors on StockTwits plays a pivotal role in diffusing information from professional to novice investors, thus diffusing firm-specific information to relevant segments of the markets.

We use the instrumental variable approach to alleviate potential endogeneity concerns from omitted variables or reverse causality and rely on two-stage least square (2SLS) regression. One of the distinct advantages of investor-oriented social media platforms is that they represent a highly diverse group of investor communities. Rubinstein (1993) suggests that the diversity of agents in the market may be attributed to their opinions and risk preferences. Therefore, we use two instruments to capture the diversity of opinions of investors on StockTwits. Following

[91] StockTwits provides three options to investors to select their investment experience, i.e., professional, intermediate and novice. Following Cookson and Niessner (2019), we regroup novice investors by putting intermediate and novice investors into one group.
[92] For the purposes of this analysis, we discard ideas posted by investors who do not disclose their investment experience.

Dow and Karunaratna (2006) and leveraging the benefits of multiple data points harvested from StockTwits, we use self-disclosed investors' geographic locations[93] to measure psychic distance (*PSYCH_DIST*) among investors on StockTwits. *PSYCH_DIST* is based on a formative index that accounts for differences in language, education, industrial development, democracy, and religion among different countries in the world. To measure the psychic distance of each investor, we use the United States as a base country.

Our second instrument is *PROXIMITY*, which is defined as the number of investors who have the same city (*PROXIMITY*) as the firm's headquarters for whom they have been updating their sentiments on StockTwits.[94] Ivković and Weisbenner (2005) present evidence that investors who invest within 250 miles of their geographic proximity earn 3.2% of additional annual returns compared to their nonlocal investments and suggest that such local bias is information-driven. Similarly, Bodnaruk (2009) highlights the role of the local information effect and presents evidence that such investors have better access and expertise to process local information. Both instruments cover the diversity of investors' opinions from local and global perspectives.

These instruments provide an independent source of exogenous variation for each endogenous regressor and meet the valid instruments criteria. Both instruments validate the relevance restriction since an increase in the number of investors from diverse backgrounds increases the diversity of opinions, and local investors' presence increases the flow of firm-specific information, thus providing a unique opportunity for investors to update their economic models based on available information. Regarding exclusion restriction, it is highly unlikely that crash risk can influence any of the instrument variables. This is because *PSYCH_DIST* is a consolidated index representing six diversity measures, and information acquisition by local investors is a natural process that specific market events may not influence. Therefore, the results from 2SLS regression further support our hypothesis and remain consistent after controlling for endogeneity.

Investors on StockTwits consume information from various information channels. It is pertinent to note that not all channels provide value-relevant information, and prior literature on crash risk suggests multiple channels associated with crash risk. Therefore, our current

[93] Investors' geographic locations are further standardized to ensure we have the country names of all locations. So, if any investor has mentioned London as her location, we use the United Kingdom. This is because Dow and Karunaratna (2006) formative index measures psychic distance between countries only.
[94] To measure this proxy, we only use investors' city-level information as disclosed on their public profile.

findings warrant further evidence to examine prominent channels which trigger sentiment oscillations on StockTwits.

The use of stock- and option-based compensation is an attractive tool that motivates managers to maximize shareholders' wealth and discourage managerial empire building (Kim et al., 2011b). Benmelech et al. (2010) present evidence that stock-based compensation motivates managers to conceal bad news at the expense of future growth opportunities. In our research setting, we are keen to explore this channel of managerial incentives to investigate the moderating effect of stock- and option-based compensation on sentiment oscillations. Following Core and Guay (2002) and Coles et al. (2006), we measure *ΔOPTIONS* and *ΔEQUITY* for the CEOs and CFOs of our sample firms. Where *ΔOPTIONS* is defined as the dollar change in the options portfolio of the CEO/CFO with a 1% increase in stock price, and *ΔEQUITY* is defined as the dollar change in the equity portfolio of the CEO/CFO with a 1% increase in stock price. It is pertinent to note that Coles et al. (2006) calculated overall managerial incentives without differentiating between the roles of the CEO and CFO. However, motivated by Jiang et al. (2010) and Kim et al. (2011b), we measure CEOs' and CFOs' managerial incentives individually.

We find evidence that an increase in the sensitivity of CFOs' options holding draws attention to the impact of sentiment oscillations to predict firm-level future crash risk. However, we do not find any significant impact of CFOs' equity portfolios' sensitivity or CEOs' equity and options portfolios. These findings are consistent with Kim et al. (2011b), who present evidence that CFOs' option incentives increase the crash risk; and Chava and Purnanandam (2010), who argue that CEOs and CFOs influence different business segments, i.e., capital structure and debt maturity choices, respectively. To further test the validity of our findings, we filter the texts of ideas on StockTwits discussing CFOs, CEOs, Options, and Equities (*see Appendix 4.3*). It is interesting to note that investors on StockTwits keenly observe CEOs' and CFOs' trading activities and discuss these with other investors on StockTwits.

These findings provide valuable insights into our analysis. First, investors follow firms' executives' activities by discussing them on StockTwits and sharing their analyses. Second, investing in options is a useful risk hedging method for executives and investors. Therefore, close monitoring of executives' options holdings on StockTwits is a natural choice for investors since such trades increase the salience of information signals. Finally, building on their analyses and learning from social interactions with peer investors, option-based compensation for managers substantiates the impact of sentiment oscillations on firm-level future crash risk.

The motivation for investors to access investor-oriented social media platforms is to get information leads to understand the correct fair value of the firm and discuss the outlook for stocks. For this purpose, investors rely on multiple information channels. We extend our analysis to explore the impact of these channels on sentiment oscillations and crash risk. First, a reliable source of firm-specific information for investors is the firm's financial statements freely available from EDGAR and/or company websites. However, it is not necessary for financial reports to be presented with complete accuracy and the highest level of accounting conservatism. For example, Hutton et al. (2009) provide evidence that firm-level opacity is positively associated with the firm's future crash risk.[95] However, it is vital to test whether investors on StockTwits can differentiate between firms based on financial reporting conservatism. Following Khan and Watts (2009), we use *CSCORE* as a firm-level proxy for accounting conservatism. We find that the impact of sentiment oscillations on future crash risk is more pronounced for firms with low conservatism, and there is no significant association between sentiment oscillations and crash risk for firms with high conservatism. These results lend credence to our previous findings and suggest that investors on StockTwits pay more attention and discuss firms with lower levels of accounting conservatism and update their sentiment accordingly.

Second, besides acquiring firm-specific information from financial statements, firms' information environment plays a pivotal role in sending information signals to all stakeholders to enrich their analyses. In this regard, firms' analyst coverage and product market competition are common indicators for the quality of monitoring and self-discipline of managers, respectively (Dyck et al., 2010; Giroud & Mueller, 2011). In our research setting, such information signals to investors on StockTwits increase (decrease) investors' confidence for firms with high (low) analysts' coverage and higher (lower) product market competition. To test this conjecture, we extend our analysis to investigate the moderating effect of analyst coverage and product market competition. We find that lower analyst coverage substantiates the ability of sentiment oscillations to predict firm-level future crash risk. These findings are consistent with prior literature (An et al., 2020; Kim et al., 2019), highlighting the facilitative role of investor-oriented social media platforms and investors on StockTwits to reduce

[95] We control for firm-level opacity in all our regressions.

information asymmetry and provide firm-specific information for firms with less analyst coverage.

Regarding product market competition,[96] we find that the impact of sentiment oscillations on firm-level future crash risk is substantiated for firms with a lower level of product market competition. These findings are consistent with prior literature (Giroud & Mueller, 2010, 2011; Kim et al., 2011b), suggesting that higher product market competition acts as a substitute for conventional governance mechanisms. Therefore, managers are less disciplined and have more motivation to conceal bad news when there is less competition. These findings complement our existing findings, further suggesting that investors on StockTwits allocate their attention to firms with weak governance mechanisms, since such firms are more prone to firm-specific future crash risk.

Third, prior literature provides corroborative evidence suggesting that aggregate market sentiment influences investors' trading behaviors in financial markets (Antoniou et al., 2013, 2016; Stambaugh et al., 2012). Motivated by these findings, we examine the role of aggregate market sentiment to influence investors' opinion on StockTwits. We use Baker and Wurgler (2006, 2007) and Huang et al. (2015) market sentiment indices and find substantial evidence that positive market sentiment increases investors' confidence in information signals and convinces investors to update their sentiment. In the case of negative market sentiment, we find weak and sparse evidence. These findings are consistent with Daniel et al. (1998), who present evidence that investors overestimate their signals' accuracy during high sentiment periods and become overconfident; and Barberis (2018), who suggest that investors' overconfidence convinces them to update their economic models based on available information.

To check the robustness of our results, we use alternative machine learning approaches, i.e., Random Forest Decision Trees (*RFDT*) and Support Vector Machines (*SVM*), to predict investors' sentiment and measure sentiment oscillations. With regard to measuring crash risk, we use three additional crash risk measures, i.e., negative coefficient of minimum return ($NCMRET_{k,t}$) which is a continuous variable, and two other measures, *CRASH320* and *CRASH20,* which are dichotomous variables. Our results remain consistent and present a positive association between sentiment oscillations and firm-level future crash risk.

[96] Product market competition is measured using Hirfindahl Hirschman's index of firms' yearly sales.

This study contributes to the existing literature on crash risk by offering substantial evidence that investors on social media platforms can predict firm-level future crash risk. We accomplish this by offering a unique proxy for updates in investors' sentiment on StockTwits, i.e., sentiment oscillations. To the best of our knowledge, this is the first study that examines the relationship between sentiment oscillations on investor-oriented social media platforms and crash risk. The research contribution of this study is threefold. First, despite the significance of crash risk, there is a lack of evidence to investigate the role played by investors in financial markets to predict firm-level future crash risk. Our study provides firsthand evidence that social interaction among investors on StockTwits facilitates these investors to consume firm-specific information, which enables them to anticipate firm-level future crash risk. This evidence is consistent with the difference of opinion theory from Hong and Stein (2003), which provides a behavioral explanation for crash risk.

Second, our findings present empirical evidence for the role of investors' heterogeneity and information diffusion via a large social network of investors in financial markets and how these investors' specific attributes enable them to predict firm-level future crash risk. It is pertinent to note that despite the critical role of the heterogeneity of investors in financial markets, as suggested by Rubinstein (1993) and Kandel and Pearson (1995), there is no direct evidence that explores the role of the heterogeneity of investors at the market level. Similarly, Hong and Stein (2003) suggest that crashes are contagious as they spread market-wide and affect other closely correlated stocks. Benefiting from an extensive social media data set, we present evidence that contagion occurs when the velocity of information diffusion is increased via a large social network of investors.

Third, this study highlights the role of investor-oriented social media platforms in financial markets. Although there is burgeoning literature investigating the role played by investor sentiment to predict volumes, returns and analyze investors' trading behaviors in financial markets, there is a lack of evidence highlighting the role of such platforms. Our study provides direct evidence on (1) how investor-oriented social media platforms assist investors to sift through an abundance of information to make investment decisions and update their analyses; (2) how social interactions and the diversity of opinion among investors attract investors, reduce information asymmetry, and rejuvenate financial markets by increasing volumes; (3) the effect of assisting investors to access information at low/no cost and providing the regulators with an opportunity to monitor financial markets meticulously. Overall, our study highlights the significance of investor-oriented social media platforms and their contribution to integrating financial markets.

Although no previous studies have investigated the role played by sentiment oscillations on social media platforms for investors seeking to predict firm-level future crash risk, our study is closely related to An et al. (2020), who highlight the monitory and disciplinary role of (*traditional*) media and provide evidence that greater media coverage discourages managers from withholding bad news. However, our study provides evidence and complements the existing literature by suggesting that traditional media and social media platforms for investors are two distinct information sources for investors. This is because traditional media are formal, well-organized, and only provide specific news and analyses without accounting for market participants' commentaries, thus supplying one-sided views to their audience. In the case of social media platforms for investors, investors share their opinions, provide feedback on news and benefit from expert opinions by learning investment analysis techniques. Therefore, traditional media pose a reputational risk for managers by discouraging them from withholding bad news, and social media for investors assist investors to predict firm-level future crash risk.

The remainder of the paper is structured as follows: Section 2 provides the institutional background and conducts a review of the literature. Section 3 presents the theoretical framework and a testable hypothesis. Section 4 presents the research design. Section 5 presents the empirical analysis. Section 6 presents robustness checks, and section 7 concludes the study.

4.2. Institutional Background and Review of Literature

4.2.1. Role of Investor Sentiment in Financial Markets

According to Antoniou et al. (2013), sentiment is defined as individuals' pessimism or optimism based on exogenous factors about a particular market event. A large body of literature suggests that sentiment affects individuals' judgment and decisions. In this line of research, the seminal work of Kahneman and Tversky offers a realistic view of how investors form beliefs and make decisions (Kahneman, 1973; Tversky & Kahneman, 1973, 1974). We define sentiment oscillations as revisions to investors' beliefs. Our measure of sentiment oscillations[97] is consistent with Kandel and Pearson (1995) definition of revisions in beliefs where they use recommendation revisions by the research analysts of a brokerage house. In a similar line of

[97] In addition to Kandel and Pearson (1995), our sentiment oscillation measure is also consistent with other studies in Economics and Finance that use analyst revisions as covariates in their studies. However, the main difference between our oscillation measure and analyst revisions is that investors on social media platforms can update their recommendations at a higher frequency compared to analysts.

research, previous studies have focused on positive and negative investor sentiment (Antweiler & Frank, 2004; Azar & Lo, 2016; Das & Chen, 2007), investors' opinions on social media platforms (Chen et al., 2014; Sprenger & Welpe, 2011), convergence and divergence of investors' opinions on social media and news (Giannini et al., 2019), and disagreements among investors (Cookson & Niessner, 2019).

Investor sentiment plays a pivotal role in understanding the various behavioral and psychological factors that drive financial markets. Previous studies starting with Markowitz (1952) and Modigliani and Miller (1958) rely on two popular assumptions about investors' trading behavior in financial markets. First, all investors in financial markets are rational, and they update their beliefs when new information arrives. Second, investment decisions by investors are driven by expected utility. However, Shiller (1981) offers an interesting view, suggesting that there are rational and less rational investors in financial markets, and their rationality affects their investment decisions. Similarly, De Bondt and Thaler (1985) present a stock market model for under- and overreaction to certain market events, further suggesting that not all market participants are rational. Besides other seminal work in behavioral finance,[98] these studies motivated subsequent research in finance to understand the interaction between rational and less rational investors and belief formation in financial markets.

To understand changes in investor sentiment, Harris and Raviv (1993) and Kandel and Pearson (1995) present differential-interpretation hypotheses, further suggesting that investors interpret information signals based on their preferences and economic models.[99] Although these papers follow very similar approaches, they differ based on the assumptions used in their models to explain changes in sentiment. For example, Kandel and Pearson (1995) assume that all investors have heterogeneous priors and are risk-averse. In contrast, Harris and Raviv (1993) assume that investors have homogenous priors. However, how investors interpret information signals varies, resulting in the differential interpretation of information signals and a change in investor sentiment. In a similar line of research, Bamber et al. (1999) argue that differential interpretation is a key driver of speculative trading in financial markets.

[98] For example, Hirshleifer (2001); Shleifer (2000) and Barberis and Thaler (2003) present various psychology-based models to explain asset prices in financial markets. Similarly, Baker and Wurgler (2013) and Malmendier (2018) provide detailed surveys of the role of behavioral corporate finance.
[99] These arguments contradict to the traditional arguments of Markowitz (1952) and Modigliani and Miller (1958), who suggest that investors follow Baye's rule to change their sentiment.

4.2.2. Firm-level Crash Risk and Investors' Sentiment

There is growing literature investigating the causes and effects of firm-level crash risk in financial markets. In this line of research, Kothari et al. (2009) argue that firm-level crash risk is linked to managers withholding bad news for a longer period but eventually being unable to control its flow. There may be several endogenous and exogenous factors that motivate managers to withhold bad news, e.g., executives' career concerns (Kim et al., 2011b), firm opacity (Hutton et al., 2009), powerful CEOs and their wealth concerns (Al Mamun et al., 2020) and stock liquidity (Chang et al., 2017) are some prominent endogenous factors. Similarly, product-market threats (Li & Zhan, 2019), the proximity to SEC head office (Kubick & Lockhart, 2016), and weakened monitoring by distracted institutional investors (Ni et al., 2020) are some prominent exogenous factors that can motivate managers to withhold bad news.

To the extent that managers can withhold bad news, recent studies have also highlighted some regulators' key steps and external stakeholders that can deter managers from hoarding bad news. In this vein, Chen et al. (2018) present evidence that a crackdown on corruption reduces the probability of firm-level crash risk since it discourages managers from withholding bad news. Similarly, An et al. (2020) provide evidence that greater media coverage deters managers from withholding bad news. They emphasize that media pose a higher reputational risk for managers and act as an external monitor to ensure a transparent information environment for all stakeholders. This view is also consistent with Dang et al. (2018), who argue that highly leveraged firms are exposed to greater monitoring by external stakeholders, thus reducing firm-level crash risk.

From an investor's perspective, it is pertinent to mention that firm-level crashes are rare events. However, such events may affect investors' judgment and decisions to predict future crash risk. Through the lens of Tversky and Kahneman (1974), they categorize such behaviors into two heuristics. First, availability heuristics suggest that individuals overestimate the occurrence of recent events and vice versa. For example, Goetzmann et al. (2017) suggest that individuals overestimate the non-occurrence of another crash after many years of crashes. Similarly, Marks (2011) and Gennaioli et al. (2012) argue that availability heuristics convince individuals to underestimate the probability of a crash, resulting in investors taking large levered positions in risky assets. However, under such circumstances, they become vulnerable to bad news, which exacerbates the probability of firm-specific crash risk. Second, representativeness heuristics convince individuals to view only positive information signals, and any analysis based on such information signals underestimates left-tail occurrences

(Bordalo et al., 2019). It is pertinent to note that investors overestimate their judgments under both heuristics, resulting in investors' overconfidence (Barberis, 2018).

A recent study by Chang et al. (2020) presents corroborative evidence using staggered EDGAR implementation that less disagreement among investors reduces crash risk. They argue that disagreement is the central idea that connects three key mechanisms, i.e., investors' overconfidence, limited attention, and information diffusion. Overall, the extant literature provides unified arguments using traditional proxies for sentiment that investors' behavior plays a pivotal role in understanding crash risk and improving information asymmetry in financial markets.

4.3. Theoretical Framework and a Testable Hypothesis

4.3.1. Theoretical Framework

The *Difference-of-opinion* theory presented by Hong and Stein (2003), henceforth Hong-Stein, provides a behavioral explanation of crash risk. We follow the Hong-Stein model to predict firm-level crashes using the sentiment oscillations of investors on StockTwits. According to the Hong-Stein model, when optimistic and pessimistic investors revise their recommendations, it results in the asymmetric distribution of returns, whereas the net distribution of returns is negatively skewed. It is pertinent to note that both types of investors face short-sale constraints, are not fully rational, and believe in their economic models in an open market; even other investors reveal their economic models.[100]

In an efficient market, when all investors receive information signals from various sources, pessimists keep out of the market because their valuation of stock prices lies below that of optimists buying the stock at current prices. Trading will only occur between optimists and unconstrained rational arbitrageurs who are risk-neutral. Since pessimists are keeping out of the market, it may not be possible to predict their stock price valuations. However, it is quite evident that their valuations lie far below those of optimists, for that reason, they are keeping out of the market. As long as optimists continue their buying spree and pessimists stay out of the market, pessimists' valuation models will remain hidden. However, when optimists change their recommendations based on available information signals and start selling, that will be when the market starts realizing the magnitude of negative information held by pessimists. In that case, the variance will be greater when prices are falling as compared to rising. Therefore,

[100] This is a deviation from the Bayesian approach and can be thought of as investors' overconfidence which results in ongoing disagreements with other investors in financial markets (Barberis, 2018).

greater variance on the downside implies an increase in the negative skewness of returns. In contrast, if pessimists do not step in and prices keep falling, this will worsen the situation and increase the crash risk. Under such circumstances, changes in investors' recommendations warrant finding further evidence to understand the implications for such revisions to predict crash risk.

It is pertinent to note that previous literature also provides theories to understand large price movements and information asymmetries in financial markets. We categorize these theories into three groups: 1) incomplete information aggregation models for rational investors, 2) volatility feedback models, and 3) behavioral models. However, the Hong-Stein crash theory has distinct empirical implications, as compared to previous studies, as follows. First, unlike Romer (1993), who explains symmetric movement in his information aggregation model for rational investors, the Hong-Stein model is asymmetric, i.e., explaining downward price movements compared to upward movements.[101] Second, like other representative agent framework models, volatility feedback models (French et al., 1987; Pindyck, 1984) ignore the role of trading volume and rely on higher inflows of news to generate large price movements. The Hong-Stein model presents the other side of the story by focusing on trading volumes as a proxy for investors' differences of opinion.[102] The Hong-Stein model complements behavioral models that discuss the implications of investors' psychological bias in financial markets and warrant seeking further evidence to unleash the role of investors' sentiment in generating negative asymmetries resulting in large price movements.

4.3.2. Testable Hypothesis

Building on the theoretical framework of Hong-Stein, we use sentiment oscillations on StockTwits to predict crash risk, where sentiment oscillations are defined as a change in investors' sentiment from *bullish-to-bearish* and vice versa. Chen et al. (2001), following the Hong-Stein theoretical framework, present that a shift in investor sentiment sets the stage for negative skewness. On similar lines, we are keen to understand the role of investor sentiment because of the following characteristics of investor-oriented social media platforms.

[101] This argument is consistent with historical large market movements. For example, the largest price gains in S&P 500 since 1923 is 16.61% on 15 March 1933. More recently, the largest price gains in one day were recorded on 24 March 2020 and 13 March 2020 as 9.38% and 9.29%, respectively. On the contrary, the largest one day decline was recorded as 20.47% on 19 Oct 1987. More recently the largest declines were recorded on 16 March 2020 and 12 March 2020 as 11.98% and 9.51%, respectively.
[102] Hong and Stein (2003) define crashes as the large shifts in stock prices in the absence of public news event.

First, investor-oriented social media platforms offer heterogeneity of investors, which when using traditional proxies[103] for sentiment is a hard-to-observe element. Kandel and Pearson (1995) argue that the heterogeneity of agents in financial markets is the most common source of differential interpretation of information signals. Therefore, in a market with heterogeneous agents, all agents have different economic models, and they update their beliefs based on those models (Hong & Stein, 2007). Since social media platforms for investors offer a unique opportunity for all investors to interact and share their opinions, such platforms have a transactional role in transmitting information representing investors' opinions.

Second, investors' large social networks are ideal for triggering contagion in financial markets.[104] This is because heterogenous investors in a large social network can highlight firms withholding bad news or releasing less information. This will increase information diffusion in a large social network and encourage investors to update their models. For example, Da et al. (2020) provide evidence of how investors form their opinions based on the most recent events, highlighting crowd wisdom in financial markets.[105] Third, Unlike previous studies that mainly rely on analysts' forecasts revisions (Bamber et al., 1999; Kandel & Pearson, 1995), our sentiment oscillation measures act as a direct proxy to account for revisions to investors' sentiment. Similarly, Karpoff (1986) and Kim and Verrecchia (1991) argue that revisions to agents' beliefs can be more informative than disagreements among investors in financial markets. Therefore, our sentiment oscillation is a unique proxy for revisions.

Following the Hong-Stein two-period model, we assume a continuum of heterogeneous investors in a large social network that provides investors with a unique opportunity to update their sentiment based on available information. At time 1, both optimist and pessimist investors analyze available information signals and, based on their interpretation, decide to take a position (optimists) or stay out of the market (pessimists). At time 2, when updated information arrives in the market, based on their analyses and interpretation of information signals, investors can stay put if updated information signals coincide with their economic models or revise their sentiment. Hong-Stein suggests that the arrival of bad news convinces optimistic investors to revise their sentiment (*bullish-to-bearish*). Similarly, pessimists do not revise their sentiment on the arrival of bad news since they are already keeping out of the market. Under such circumstances, revisions to investors' sentiment should predict the negative skewness of

[103] For example, Volume, Price, Analyst dispersion and Market Volatility are commonly used proxies.
[104] According to the Hong-Stein theory of crashes, a shift in stock prices not only affects a single stock but also other stocks whose returns are positively correlated to falling stocks. Therefore, crash risk is contagious.
[105] For further explanation of the wisdom of crowds, see Surowiecki (2005) and Chen et al. (2014).

returns as more information dribbles out, which coincides with pessimist investors' economic models. Therefore, our hypothesis is as follows:

H1: *Ceterus Paribus, firms with greater sentiment oscillations are associated with higher levels of future stock price crash risk.*

4.4. Research Design

4.4.1. Sample Selection and Variables Measurement

To examine the relationship between investors' opinions on StockTwits and crash risk, we include all U.S. firms listed in NYSE/NASDQA/AMEX with share codes 10 and 11, from Jan. 2012 to Dec. 2017. Stock price data are collected from the Center for Research in Security Prices (CRSP). Firm-level financial statement data are collected from Compustat, IBES, and Execucomp databases.

Our social media data are collected from StockTwits, which is a popular social media platform for investors. StockTwits is by far the largest social media community for investors and traders, with more than three million members, five million monthly messages, and three million monthly visitors.[106] The main user interface of StockTwits is user-friendly; it is where investors can post Twitter-like messages with a 140-character limit.[107] One of the distinguishing features of StockTwits is investors' user profiles, where any investor can volunteer to disclose their asset choices, investment approaches, and investment term preferences. Moreover, investors can create a customized watchlist to view StockTwits' ideas directly relevant to their investment preferences. To collect data from StockTwits, we developed a python program to connect with StockTwits API and collect multiple data points based on numerous iterations from Jan. 2012 to Dec. 2018. StockTwits relies on cashtags ($AAPL) as company identifiers. We were able to harvest more than 38 million ideas during our data collection, posted by approximately 297,000 users, discussing 8,394 companies listed in the U.S. and the rest of the world's stock exchanges.

We apply several filters to ensure the quality of data harvested from StockTwits. To avoid concerns related to shared attention and discussion of multiple cashtags in a single message, we only keep StockTwits' ideas that contain single cashtags and only discuss one

[106] https://about.stocktwits.com/

[107] On 8 May 2019, StockTwits increased their character limit to 1,000 characters. However, this is outside our sample period.

company in a given message. Our final filter restricts our StockTwits sample to common stocks. This filter is essential in terms of investors' ability to identify companies and discuss their ideas uniquely. After applying these filters, we are left with approximately 20 million ideas containing 2,270 unique cashtags. We match the StockTwits sample with our U.S. firm-level data, resulting in 1,815 firms and 104,137 firm-month observations in our final sample, with more than 13 million StockTwits ideas.

4.4.2. Measuring Sentiment Oscillations

StockTwits is a dedicated investor-oriented platform where any investor can voluntarily disclose their recommendations. However, not all investors disclose these. Therefore, we use the Maximum Entropy (MaxEnt) model to predict the recommendations of investors who did not disclose theirs in the first place. MaxEnt is derived from the exponential class of models and follows the maximum entropy approach, i.e., it selects the model with the highest entropy from all models that fit our training data. One of the key advantages of using the MaxEnt model is its minimum assumptions compared to the Naïve Bayes models. MaxEnt is commonly used for text analysis and when the prior distribution of data is not known.

In a next step, we train StockTwits data to predict investors' recommendations. In this regard, training data size is a critical factor in determining overall prediction accuracy. For this purpose, we use 1.2 million self-labeled ideas by investors and use these as a training data set. To test the post-training classification's validity, we use a 10-fold cross-validation test representing 77% of the area under the curve (AUC Score), the prediction accuracy of 74%, and an F1 score of 82%. Our final test to check our prediction model's validity is to use feature selection to train the data set. This is because, in a large data set, some less relevant features may impact on prediction models' performance. We use automatic feature selection because it can train the model (comparatively) faster, improve a complex data set’s prediction accuracy, and reduce the overfitting of models. Our results remain consistent after using feature selection.

In a final step, we measure sentiment oscillations (*SENT.OSC*). For each sentiment oscillation, we assign a weight based on an investor’s number of followers. Assigning weights to sentiment oscillations allows us to address two key factors. First, using the number of followers, we account for influential investors within the network. Therefore, influential investors whose sentiment oscillations may have more impact than less influential investors will have greater weight. Second, a large social network represents information that can be diffused at a higher velocity within the network. Considering the contagion effect suggested by

Hong and Stein (2003), we assume that a large social network has the clear ability to create information contagion that can exacerbate the spread of information within a large social network.

To measure *SENT.OSC*, following Kwak et al. (2010), we define a set of heterogeneous investors on StockTwits by a directed graph as follows:

$$G = (V, E) \tag{1}$$

V is a set of vertices uniquely identifying each investor. E is a set of ordered edges, which is a collection of directed pairwise social media connections between investors on StockTwits. This is equivalent to an investor having several followers (inward connections) and several other investors they follow (outward connections). Each investor can post their recommendations (*Bullish* or *Bearish*) on the platform, and those recommendations will be distributed over the network according to the connections in graph G. Each investor (i.e., each vertex V) has a certain number of inward connections and outward connections in the graph. The number of inward connections is known as the indegree of the vertex. Since we are only interested in the number of followers to calculate the influence of each investor within the network, we only account for the in-degree network of the vertex as $deg_{in}(V)$. We assign weights to investors' recommendations as follows:

$$\omega_i = \frac{deg_{in}(V_i)}{|E|} \tag{2}$$

Where ω_i is the weight assigned to investor i, $deg_{in}(V_i)$ is the in-degree network of investor i. Meanwhile $|E|$, is defined as follows:

$$|E| = \sum_{i=1}^{N} deg_{in}\,(V_i) \tag{3}$$

Therefore, $|E|$ is the total number of inward edges of investor i, in graph G. We define investors' recommendations as follows:

$$\rho_{k,t}(V_i) \ \in [0,1] \tag{4}$$

Where $\rho_{k,t}(V_i)$ is the recommendation of unique investor i, for a set of stocks k at time t, and 0 and 1 indicate bearish and bullish recommendations, respectively. Each investor may change their recommendations over time as they receive more information or change their prior beliefs about a stock. Thus, we define investor V_i's revision of their recommendation for stock k from $t - 1$ to t as the absolute value of the difference between the two recommendations and call these revisions sentiment oscillations, hence:

$$\delta_{k,t}(V_i) = |\rho_{k,t}(V_i) - \rho_{k,t-1}(V_i)| \tag{5}$$

Where $\delta_{k,t}(V_i)$ is the sentiment oscillation for each investor (V_i) of stock k at time t. The value of this sentiment oscillation will be equal to 1 if they change their recommendation from *bullish-to-bearish* or vice versa, or it will be equal to 0 if their recommendation remains the same. We can then count the total number of sentiment oscillations at time t for different stocks in the sample by each investor as follows:

$$\Gamma_t(V_i) = \sum_{k=1}^{K} \delta_{k,t}(V_i) \tag{6}$$

Where $\Gamma_t(V_i)$ is the total sentiment oscillations by each investor (V_i) in the network for different stocks in our sample. For each stock k we calculate for each month m the monthly weighted aggregate of investor V_i's oscillations as:

$$\Pi_{k,m}(V_i) = \omega_i \sum_{t=1}^{d_m} \delta_{k,t}(V_i) \tag{7}$$

Where $\Pi_{k,m}(V_i)$ is the total number of weighted sentiment oscillations (*SENT.OSC*) for each stock k in month m, and ω_i is the weight assigned to each investor in Equation (2). Our final measurement of *SENT.OSC* is the natural logarithm of the three months' moving average of sentiment oscillations. We use a three months moving average to preclude the impact of any exogenous shocks on investors' recommendations.

$$SENT.OSC_{k,t} = \ln\left[1 + \left(\frac{1}{3}\sum_{n=m}^{m+3} \Pi_{k,m}(V_i)\right)\right] \tag{8}$$

4.4.3. Measuring Crash Risk

We use daily stock returns data to calculate monthly firm-specific crash risk measures. We follow Chen et al. (2001); Hutton et al. (2009); Kim et al. (2011a); Kim et al. (2011b) and Baloria and Heese (2018) to filter our data set by excluding financial firms (6000 ≤ SIC ≤ 6999) and utility firms (4000 ≤ SIC ≤ 4999) and stock values of less than $1. We exclude firms with less than 26 weeks of data to ensure consistency with previous studies (Kim et al., 2011a; Kim et al., 2011b; Kim & Zhang, 2016). Motivated by the variety of crash risk measures used in recent literature, we employ six different monthly firm-specific crash risk measures, i.e., three continuous and three dichotomous variables. We use $NCSKEW_{k,t}$, $DUVOL_{k,t}$ and *CRASH309* as our primary measures of crash risk in the analysis and use $NCMRET_{k,t}$, *CRASH320* and *CRASH20* as alternative measures of crash risk to check the robustness of our results. We use alternative crash risk measures to address the limitations of our primary crash risk measures.

The first step in calculating firm-specific crash risk is to remove the impact of market returns. Therefore, to obtain firm-specific returns, we estimate the following regression based on a 52-week rolling window starting from week 0:

$$R_{k,w} = \alpha_k + \beta_{1,k}.R_{m,w-2} + \beta_{2,k}.R_{m,w-1} + \beta_{3,k}.R_{m,w} + \beta_{4,k}.R_{m,w+1} + \beta_{5,k}.R_{m,w+2} + \varepsilon_{k,w} \tag{9}$$

Where $R_{k,w}$ is the weekly return on stock *k* in week *w* and $R_{m,w}$ is the weekly market return in week *w*. We use the value-weighted CRSP return index as market return. Following Dimson (1979), we add lead and lag market returns to deal with the nonsynchronous trading issue for less frequently traded stocks. From Equation (9), we calculate firm-specific weekly returns as the log of residual returns as follows:

$$W_{k,w} = \ln\left(1 + \widehat{\varepsilon_{k,w}}\right) \tag{10}$$

The first measure of crash risk is the negative conditional skewness ($NCSKEW_{k,t}$) of firm-specific weekly returns. Following Chen et al. (2001), we calculate $NCSKEW_{k,t}$ by taking the third moment of $W_{k,w}$ and then dividing it by the standard deviation of $W_{k,w}$ raised to the third power as follows:

$$NCSKEW_{k,t} = -\left[n(n-1)^{\frac{3}{2}} \sum W_{k,w}^{3}\right] \Big/ \left[(n-1)(n-2)\left(\sum W_{k,w}^{2}\right)^{\frac{3}{2}}\right] \tag{11}$$

Where $W_{k,w}$ is the firm-specific weekly returns for firm k in week w, and n is the number of firm-specific weekly returns calculated in Equation (10) based on a 52-week rolling regression. The higher the value of $NCSKEW_{k,t}$ the greater the negative conditional skewness of the returns and the higher the crash risk. Although $NCSKEW_{k,t}$ is a commonly used crash risk measure, there are two limitations of negative conditional skewness of returns. First, firm-level crashes are defined as large negative returns. However, negative skewness can also occur due to a large number of less extreme negative returns. Second, this measure does not cover both crashes and jumps due to abrupt changes in stock prices (Ak et al., 2016). We address these limitations by constructing alternative measures of crash risk.

The second measure of crash risk is the degree of up/down volatility ($DUVOL_{k,t}$), which is calculated as follows:

$$DUVOL_{k,t} = log\left[\frac{(n_U - 1)\ \sum_D W_{k,w}^2}{(n_D - 1)\ \sum_U W_{k,w}^2}\right] \tag{12}$$

Where $DUVOL_{k,t}$ is the degree of up/down volatility of firm k at time t, n_U is the number of up weeks at time t and n_D is the number of down weeks at time t. An up (down) week is a week when the firm-specific weekly return $W_{k,w}$ is greater (less) than the average of firm-specific weekly returns estimated based on a 52-week rolling regression. The higher the value of $DUVOL_{k,t}$, the higher the crash risk of the firm.

The third measure of crash risk (*CRASH309*) is a dichotomous variable that equals 1 if the firm-specific weekly return decline more than 3.09 standard deviations below the average firm-specific weekly return within a given month, where *CRASH* is defined based on the 52-week rolling regression estimated in Equation (9). Following Andreou et al. (2017); Kim et al. (2011b), we estimate *CRASH320* as the fourth measure of crash risk based on firm-specific weekly return decline more than 3.20 standard deviations below the average firm-specific weekly return within the given month and create a dichotomous variable that equals 1 under such circumstances. *CRASH* is defined based on the 52-week rolling regression estimated in Equation (9). Motivated by Andreou et al. (2021), we estimate a model-free crash risk measure, i.e., *CRASH20*. *CRASH20* is calculated as a dichotomous variable that equals 1 if the market-adjusted weekly return falls by more than 20% within a given month and 0 otherwise. Market-adjusted weekly returns are calculated as the firm's weekly returns minus value-weighted weekly returns from CRSP. One of the main limitations of these dichotomous variables is that such binary variables do not indicate the magnitude of the crash.

We try to address these limitations in our final crash risk measure, i.e., the negative coefficient of minimum return ($NCMRET_{k,t}$). Following Ak et al. (2016) and Andreou et al. (2021), we define $NCMRET_{k,t}$ as the negative ratio of minimum weekly returns within the last 26 weeks divided by the sample standard deviation of the returns in the previous period.

$$NCMRET_{k,t} = -\left[\frac{\min\left(R_{k,w}, R_{k,w-1}, R_{k,w-2}, \ldots, R_{k,w-n+1}\right)}{\sqrt{\frac{\sum R_k^2}{(n-1)}}}\right] \tag{13}$$

Where $R_{k,w}$ is firm k's market-adjusted stock returns in week w over 26 weeks including the current month. Market-adjusted returns are measured as the difference between weekly stock returns and value-weighted weekly returns from CRSP. The same as in Equation (11), we use the minus sign to highlight the magnitude of left-tail skewness, i.e., the higher the value of $NCMRET_{k,t}$, the greater the crash risk.

4.5. Baseline Model

To examine the relationship between sentiment oscillations ($SENT.OSC_{k,t}$) and crash risk, we estimate the following regression equation:

$$Crash\ Risk_{k,t} = \beta_0 + \beta_1 . SENT.OSC_{k,t-1} + \gamma CONTROLS_{k,t-1} + \varepsilon_{k,t} \tag{14}$$

Where $Crash\ Risk_{k,t}$ represents one of six crash risk measures. In our main analysis, we use $NCSKEW_{k,t}$, $DUVOL_{k,t}$ and *CRASH309* in our primary analysis and, for robustness checks, we use $NCMRET_{k,t}$ *CRASH320* and *CRASH20* as alternative measures of crash risk. $SENT.OSC_{k,t-1}$ is the one-month lagged value of sentiment oscillations from Equation (8). Depending on the crash risk measure, we use ordinary least square regression with fixed effects for continuous dependent variables and logit regression with fixed effects for binary dependent variables. To deal with firm-level heterogeneity, we use firm fixed effects. Similarly, to account for time-varying and industry-specific movements, we also control for time and industry fixed effects. It is pertinent to note that we use firm and industry fixed effects alternatively along with time fixed effects in our baseline model and use industry and time fixed effects in all the regressions afterwards, unless otherwise mentioned. To determined industry fixed effects, we use two-digit standard industrial classification (*SIC*) codes. Finally, to account for within-group correlation that may influence standard errors, we cluster firm-level standard errors (Petersen, 2009). All continuous variables are winsorized at the 1st and 99th percentiles.

$CONTROLS_{k,t-1}$in Equation (14) represents a set of various firm-level and market-level controls. We control for industry-level sentiment oscillations $SENT.OSC_{t-1}$. Following Chen et al. (2001); Hutton et al. (2009); Kim et al. (2011a); Kim et al. (2011b), in addition to other firm-level control variables, we also control for $DTURN_{t-1}$ defined as the average monthly turnover in the last six months normalized by the average turnover in the previous 18 months, $SIGMA_{t-1}$, as the standard deviation of returns. RET_{t-1} is measured as average monthly returns, $KURT_{t-1}$ is measured as the kurtosis of firm-specific weekly returns within the given month. Following An et al. (2020), we control for media coverage ($MEDIA_{t-1}$), which is defined as the log of monthly firm-specific news articles discussing the sample firms. ROA_{t-1} is measured as the quarterly net income at the firm level normalized by total assets. BTM_{t-1} is measured as the quarterly book value of the common equity of the firm normalized by market capitalization. LEV_{t-1} is measured as the total quarterly liabilities of the firm normalized by the firm's total assets. $OPACITY_{t-1}$ is measured as in Hutton et al. (2009) as the 36-month moving sum of the absolute value of discretionary accruals. Discretionary accruals are calculated using the modified Jones (1991) model. $SIZE_{t-1}$ is measured as the log of quarterly firms' sales. Following Andreou et al. (2021) we control for goodwill ($GOOD_{t-1}$) defined as the quarterly value of the goodwill of the firm normalised by total assets, firm-level quarterly research and development expenditure normalized total assets (R_{t-1},), and to control for the difference between young and old firms we use *AGE10* as an indicator variable that equals 1 for a firm less than ten years old, otherwise 0.

4.6. Empirical Analysis

4.6.1. Descriptive Statistics

Table 4.1 presents summary statistics for all the variables used in our analysis. It is pertinent to note that our sample period is distinct from previous studies foe the following reasons. First, due to the constraints on data availability on StockTwits, our sample period is shorter[108] (2012–2017). Second, unlike previous studies, our crash risk and sentiment oscillation measures are calculated monthly, similar to Andreou et al. (2021). Panel A presents summary statistics for six crash risk measures. Overall, the summary statistics for crash risk measures are comparable to Kim et al. (2011a); Kim et al. (2011b), Callen and Fang (2015b) and Andreou et al. (2021). The means of $NCSKEW_{k,t}$, $DUVOL_{k,t}$ and $NCMRET_{k,t}$ are 0.075, 0.89 and 2.32, respectively. There are 5.6% of crashes that fall below 3.09 standard deviations and 4.5% of crashes that fall below 3.20 standard deviations. Finally, there are 1.8% of crashes where market-adjusted returns fall by more than 20%.

[*Insert Table 4.1 here*]

Panel B presents summary statistics for sentiment oscillation measures. The mean value of overall sentiment oscillations, $SENT.OSC_{t-1}$, is 1.762, compared to the sentiment oscillations by professional (1.078) and novice investors (1.173). It is pertinent to note that there are more novice investors than professional investors in our data set. The industry average for sentiment oscillations is 12.562. Panel C presents summary statistics for our baseline control variables. Besides using traditional control variables already used in prior literature (Callen & Fang, 2015b; Hutton et al., 2009; Kim et al., 2011a), we control for media coverage ($MEDIA_{t-1}$) following An et al. (2020), and control for firm-level goodwill ($GOOD_{t-1}$), research and development expenses (RD_{t-1}) and age of the firm (*AGE10*) following Andreou et al. (2021). The summary statistics of our baseline controls resemble previous studies' summary statistics in a similar area of interest (An et al., 2020; Andreou et al., 2021; Callen & Fang, 2015b; Hutton et al., 2009; Kim et al., 2011b). Overall, the standard deviation of these variables suggests wide variation in our sample. Finally, Panel D present summary statistics for the additional variables used in our further analysis. *ΔOPTIONS* and *ΔEQUITY* measures for both CEO and CFO are reported in thousands of US dollars. Their calculation is explained later in the forthcoming analysis. *CSCORE* is a proxy for firm-level financial reporting conservatism; *ANALYSTS* is the natural log of the number of analysts following the sample firms. *COMPETITION* is the sample firms' product market competition. Following Baker and

[108] We do not consider it a caveat of our study. However, prior studies(Callen & Fang, 2015b; Kim et al., 2011a; Kim et al., 2011b; Kim et al., 2019) have sample sizes greater than ten years, on average.

Wurgler (2006, 2007) and Huang et al. (2015), we use $BW.SENT_t$, $BW.SENT\_Orth_t$ and $PLS.SENT_t$, respectively, as proxies for aggregate market sentiment.[109]

Table 4.2, Panel A presents the correlation coefficient between our crash risk measure ($NCSKEW_t$) and all other variables of interest. The coefficient of correlation between crash risk ($NCSKEW_t$) and sentiment oscillations ($SENT.OSC_{t-1}$) is positive and statistically significant (0.021***). Similarly, there is a positive and statistically significant correlation between crash risk and other key control variables such as $DTURN_{t-1}$ (Chen et al., 2001) and $OPACITY_{t-1}$ (Hutton et al., 2009), and a negative correlation (-0.008**) between crash risk and $MEDIA_{t-1}$ (An et al., 2020).

Table 4.2, Panel B presents the correlation coefficients for our six crash risk measures. We have three continuous measures of crash risk and three dichotomous measures of crash risk. It is evident in Panel B that there is a high correlation between continuous measures and dichotomous crash risk measures. For example, the coefficient of correlation between $NCSKEW_t$ and $DUVOL_t$ is 0.946***, which is qualitatively consistent with prior studies (Callen & Fang, 2015b; Chen et al., 2001; Chen et al., 2018). Similarly, the coefficient of correlation between *CRASH309* and *CRASH320* is 0.873***, which is expected, because both dichotomous measures capture extreme negative returns in the market (Andreou et al., 2021). Overall, the correlation coefficients in Panel B suggest that our crash risk measures capture extreme negative returns (dichotomous measures) and the magnitude (continuous measures) of the negative skewness of extreme returns.

[*Insert Table 4.2 here*]

4.6.2. Sentiment Oscillations and Crash Risk

Table 4.3 presents ordinary least square (OLS) and logit regression results estimated based on Equation (14). We use *two* continuous ($NCSKEW_t$, $DUVL_t$) and one dichotomous (*CRASH309*) variable as proxies for crash risk. To further address the concerns of omitted variables at the firm and industry levels, we examine the sensitivity of results based on firm and industry fixed effects. Table 4.3, columns (1), (3), and (5), present results using firm fixed effects, and columns (2), (4), and (6) present results using industry fixed effects.

[109] The Baker-Wurgler data set of market sentiment is downloaded from the Jeffrey Wurgler website. The data set for PLS sentiment is downloaded from the online appendix of Huang et al. (2015), available at the Review of Financial Studies website.

The results in Table 4.3 column (2) present that one standard deviation increase (1.336) in sentiment oscillations results in a 0.040 standard deviation increase in crash risk ($NCSKEW_t$) in the following month. To further understand the economic significance of these results, we compare the coefficient of sentiment oscillations with the unconditional mean value of crash risk ($NCSKEW_t$). The comparison suggests that the coefficient of sentiment oscillations amounts to 53.33% of the value of the unconditional mean of crash risk, which is economically significant. We get qualitatively similar results when using $DUVOL_t$ as another (continuous) proxy for crash risk.

[*Insert Table 4.3 here*]

Moreover, we use logit regression and control for firm and industry fixed effects in columns (5) and (6), respectively. We calculate the marginal effects of sentiment oscillations on crash risk (*CRASH309*) from column (6) as the mean values of all explanatory variables. The magnitude of marginal effects suggests that one standard deviation increase in sentiment oscillation results in a 0.28% increase in crash risk probability. Given that the mean value of *CRASH309* in our sample is 0.056, the impact is economically significant.[110] Therefore, our results are both economically and statistically significant.

The coefficient estimates of our control variables are consistent with prior studies such as Callen and Fang (2015a); Kim et al. (2011a); Kim et al. (2011b), and Chang et al. (2017), among others. Considering that industry-level information may also influence investors' sentiment on StockTwits, we control for sentiment oscillations at the industry level[111] ($SENT.OSC\_Ind_{t-1}$). Following An et al. (2020), we control for the effect of media coverage ($MEDIA_{t-1}$). To control for the effect of firm information environment, we follow Hutton et al. (2009) and use firm-level opacity ($OPACITY_{t-1}$). The results for $OPACITY_{t-1}$ are consistent with Hutton et al. (2009). Chen et al. (2001) suggest that changes in turnover represent differences of opinion in financial markets. To account for this, we control for detrended turnover in all our regressions. However, it is pertinent to note that our primary variable of interest ($SENT.OSC_{t-1}$) captures differences in opinion more subtly than trading volume. The partial association of sentiment oscillations ($SENT.OSC_{t-1}$) and detrended turnover

[110] The magnitude of marginal effects is (0.0028/0.056*100) 5% when compared to the unconditional value of the mean of *CRASH309*.

[111] Our results remain consistent after excluding sentiment oscillations at the industry level.

($DTURN_{t-1}$) is also evident in the magnitude of their correlation coefficient which is 0.129*** as compared to 0.007** with crash risk ($NCSKEW_t$).

These results are consistent with our hypothesis, suggesting that firms with higher sentiment oscillations are associated with higher crash risk in the future. These findings highlight the transactional role of investor-oriented social media platforms as follows. First, investors on StockTwits update their economic models, and such updates can transmit meaningful information signals to predict future crashes in financial markets. Second, according to Hong and Stein (2003), crashes in financial markets occur in contagion and affect highly correlated stocks. Therefore, a large social network of investors on StockTwits can amplify the "*Contagion effect*" in financial markets.

4.6.3. The impact of Sentiment Oscillations during quarterly earnings announcements

One of our main concerns is the impact of unexpected market events that can cause a crash risk and affect the association between sentiment oscillations and crash risk. Along similar lines to An et al. (2020) and Andreou et al. (2021), we extend our analysis to investigate further the role of sentiment oscillations during quarterly earnings announcements. Kothari et al. (2009) present evidence using dividend announcements that managers successfully hide bad news via such announcements. To this end, we divide our sample into earning announcement months (*QEA*) and non-earning announcement months (*Non-QEA*) and then calculate the crash risk for *QEA* and *Non-QEA* months. We estimate regressions using our three crash risk measures used in the primary analysis for each group. We use OLS and Logit regressions for continuous and dichotomous dependent variables, respectively. Table 4.4 presents the regression results for *QEA* and *Non-QEA* samples. Overall, these results show a stronger association between sentiment oscillations and crash risk during earning announcement months compared to non-earning announcement months. For example, the results in column (1) show that one standard deviation increase in sentiment oscillations results in a 0.065 standard deviation increase in crash risk in the following month. It is pertinent to note that the coefficient of sentiment oscillations in column (2) is statistically significant but with less magnitude (0.027***).

[*Insert Table 4.4 here*]

These results suggest that managers successfully hide bad news during earnings announcement periods (Kothari et al., 2009). To understand whether our results are not biased by unexpected events, we find that investors on StockTwits predict crash risk during earnings announcement months. These results highlight the critical role of investor-oriented social media platforms to allow investors to share their opinions with other investors and learn from their peers in financial markets. Our results are consistent with An et al. (2020), who suggest that measuring crash risk during and after earnings announcement months can mitigate the impact of unexpected market events to predict crash risk. These results are also consistent with Hutton et al. (2009) and Kothari et al. (2009), who suggest that negative information held by the managers is more likely to be reflected in their earnings announcements. We extend this evidence by suggesting that investors on StockTwits closely follow such leads and predict crash risk.

4.6.4. Investors' heterogeneity on StockTwits

Investors' heterogeneity in financial markets is a critical factor that can exacerbate the impact of differences of opinion in financial markets (Kandel & Pearson, 1995). In this line of research, Hong and Stein (1999) highlight the role of market segmentation and specializations by suggesting that some investors in financial markets act as frontrunners to access value-relevant information. As a result, such investors update their valuation models earlier than others who receive the same information but with some delay. Despite the research worthiness of investors' heterogeneity in financial markets, there is no distinct examination of differential interpretations by professional and novice investors in financial markets. Therefore, in this section, we examine the impact of professional and novice investors' sentiment oscillations on predicting future crash risk. Investors are categorized based on their self-disclosed investment experience on StockTwits.[112]

For the purposes of this analysis, we discard ideas posted by investors who not disclose their investment experience. Overall, there are 51,400 investors who disclose their investment experience and post more than 6.8 million ideas on StockTwits while discussing the sample firms. Out of all investors, 18.97% are professionals and 81.03% are novices.[113] We re-estimate our regression model in Equation (14) and use the sentiment oscillations of professional ($SENT.OSC\_Pro_{t-1}$) and novice ($SENT.OSC\_Novice_{t-1}$) investors as our primary variables of interest. We use OLS and Logit regressions for continuous and dichotomous dependent variables, respectively. Columns (1), (4), and (7) present standalone coefficients of the sentiment oscillations of professional investors ($SENT.OSC\_Pro_{t-1}$), which are statistically

[112] To further ensure we clearly differentiate between professional and novice investors, we group investors with self-disclosed intermediate experience with novice investors.
[113] For this analysis we consider investors with intermediate experience as novice investors.

significant at 1%, suggesting that one standard deviation increase in professional investors' sentiment oscillations results in a 0.037 standard deviation increase in crash risk in the following month (as in column (1)). The magnitude of the coefficient of $SENT.OSC\_Pro_{t-1}$ accounts for 49.33% of the unconditional mean value of crash risk ($NCSKEW_t$). Columns (2), (5), and (8) present the standalone coefficients of sentiment oscillations of novice investors ($SENT.OSC\_Novice_{t-1}$), which are statistically significant at 1% (Columns (2) and (8)) and 10% (Column (5)), suggesting that one standard deviation increase in novice investors' sentiment oscillations results in a 0.021 standard deviation increase in crash risk in the following month (as in column (2)). Similarly, the magnitude of the coefficient of $SENT.OSC\_Novice_{t-1}$ accounts for only 28% of the unconditional mean value of crash risk ($NCSKEW_t$).

[*Insert Table 4.5 here*]

Besides examining the standalone impact of the sentiment oscillations of professional and novice investors, we further examine the relationships between sentiment oscillations and crash risk in the presence of both professional and novice investors. Columns (3), (6), and (9) present regression results using sentiment oscillations from both investor groups. It is interesting to note that the sentiment oscillations from professional investors are statistically significant in all the regressions. However, the sentiment oscillations from novice investors lose their significance in the presence of professional investors. It is pertinent to note that although novice investors' mean sentiment oscillations are higher than those of professional investors (see Table 4.1), professional investors dominate the market by influencing investors' opinions on StockTwits (Hong & Stein, 2007).

These findings provide critical insights into the role played by investors' heterogeneity on investor-oriented social media platforms to predict future crash risk in financial markets. First, although professional investors take the lead, the social interactions among professional and novice investors result in learning from each other's social interactions and predicting future crash risk. Second, investors' heterogeneity is a crucial ingredient to attract investors to interact with each other and benefit from their social interactions by updating economic models based on their learning. Third, in a wide social network of investors, such updates in economic models diffuse information signals at a higher velocity, thus providing reliable information to market participants to predict future crash risk. Our findings are consistent with the theoretical models of Hong and Stein (1999, 2007).

Thus far, the results for earnings announcements and investors' heterogeneity provide compelling evidence that an increase in sentiment oscillations on StockTwits offers valuable insights for market participants. Moreover, these results remain consistent during earnings announcements and in the presence of heterogeneous investors on StockTwits, further supporting our hypothesis.

4.6.5. Two-stage least square (2SLS) regression

Although we control for omitted variable bias by using fixed effects and additional controls to ensure our results' validity, these results may be subject to potential endogeneity concerns. We rely on the instrumental variable approach using two-stage least square (2SLS) regression to address these concerns. Our analysis is based on two instrument variables.

It is pertinent to note that investor-oriented social media platforms are not restricted to a single geographic location or a specific market segment. The investors on such platforms have diverse geographic backgrounds and cultures worldwide, representing a highly diverse group of investors sharing their opinions and following influential investors in their fields. Therefore, our first instrument variable captures the diversity of opinions on StockTwits from around the world. Following Dow and Karunaratna (2006) and leveraging the benefit of multiple data points harvested from StockTwits, we use self-disclosed investors' geographic locations to measure the psychic distance (*PSYCH_DIST*) between investors on StockTwits. *PSYCH_DIST* is based on a formative index that accounts for differences in language, education, industrial development, democracy, and religion among different countries in the world. To measure the psychic distance of each investor, we use the United States as a base country. We calculated the aggregated value of *PSYCH_DIST* and divided it by the total number of investors for each firm-month. The value of *PSYCH_DIST* ranges from 0–10, where 0 represents no difference and 10 represents the maximum difference.[114]

The second instrument measures the number of investors who have the same city (*PROXIMITY*) as the firm's headquarters for which they have been updating their sentiments on StockTwits. Ivković and Weisbenner (2005) present evidence that investors who invest within 250 miles of their geographic proximity earn 3.2% of additional annual returns as compared to their nonlocal investments and suggest that such local bias is information-driven. Similarly, Bodnaruk (2009) highlights the role of the local information effect and presents

[114] The value of *PSYCH_DIST* for all the US investors is 0 since US is the base country.

evidence that such investors have better access and expertise to process local information. Therefore, the higher the value of *PROXIMITY*, the greater the local information effect.

[*Insert Table 4.6 here*]

Therefore, *PROXIMITY* covers the diversity of investors within the United States, and *PSYCH_DIST* covers the diversity of investors from the rest of the world.

The instrumental variable approach relies on two main assumptions. First, there should be an independent distribution of excluded instruments' standard errors; and second, excluded instruments are highly correlated with endogenous regressors. Both our instruments, i.e., *PSYCH_DIST* and *PROXIMITY,* satisfy the good instrument criteria as follows. First, *PSYCH_DIST* captures the diversity of investors' opinions from the rest of the world. The higher the value of *PSYCH_DIST,* the greater the diversity of opinions from different geographic and demographic backgrounds. Such diversity of opinion will encourage investors to participate in discussions and update their sentiment based on available information. Second, *PROXIMITY* captures the diversity of opinions from local investors' perspectives. The higher the value of *PROXIMITY,* the greater the local information effect. Both instruments can only predict crash risk via sentiment oscillations. Table 4.6 presents the results of 2SLS regression. The results from the first-stage regression present a positive association between sentiment oscillations and two instrument variables.

We check the validity of our instruments based on the following tests. First, Stock and Yogo (2005) argue that weak instruments provide biased instrumental estimators. A rule of thumb for the *F-statistic* associated with first stage regression is that it should be greater than 10 (Bound et al., 1995; Staiger & Stock, 1994). The high value of *F-statistic* suggests the greater explanatory power of our instruments, and the instruments are sufficiently strong to justify making inferences from these results. The value of *F-statistic* is greater than 10 in our first stage regression. Second, the Sargan test is a test for overidentifying restrictions and tests for the exclusion condition. A *P-value* for a Sargan test higher than 0.05 suggests that the excluded instruments are correctly excluded from the estimated equation. The *P-value* of Sargan from second-stage regression is 0.831 and 0.388 in columns (1) and (2), respectively. Third, for the relevance condition, the additional results from an Anderson-Rubin test and a Kleibergen-paap test reject the null hypothesis (*P-values* smaller than 0.05), suggesting that the model is identified. Therefore, the association between endogenous regressors and instrument variables is adequate to identify the equation.

4.6.6. Triggers of Sentiment Oscillations on StockTwits

Thus far we have provided corroborative evidence that sentiment oscillations on StockTwits can predict crash risk. In this section, we further investigate multiple channels that trigger sentiment oscillations on StockTwits.

4.6.6.1. Managerial Investment Incentives and Bad News Hoarding

The use of stock- and option-based compensation is an attractive tool to motivate managers to maximize shareholders' wealth and discourage managerial empire-building (Kim et al., 2011b). However, in the recent past, corporations have overindulged themselves in this type of compensation, consequently attracting media and researchers to investigate why corporate executives and boards make such decisions. In this regard, Bebchuk (2009) provided a written testimony by arguing that stock options holdings encourage managers to get involved in short-termist behavior and inflate stock prices instead of preferring firms' value creation in the long run. Similarly, Benmelech et al. (2010) present further evidence suggesting that stock-based compensation motivates managers to conceal bad news at the expense of future growth opportunities. Recent studies have disintegrated various roles of managers further to understand their influence in bad news hoarding. For example, Jiang et al. (2010) show that CFO equity incentives play a pivotal in determining firms' earnings management, and Chava and Purnanandam (2010) suggest that CEO and CFO incentives are directly associated with capital structure and debt maturity choices, respectively.

In our research setting, we are keen to understand the influence of managerial incentives on investors on StockTwits to update their economic models. We expect that an increase in stock- and option-based incentives should lure investors into investigating these firms' stocks by analyzing their financial performance and interacting with other investors on StockTwits. Following Core and Guay (2002) and Coles et al. (2006), we measure *ΔOPTIONS* and *ΔEQUITY* for the CEOs and CFOs of our sample firms, where *ΔOPTIONS* is defined as the dollar change in the options portfolios of CEOs/CFOs with a 1% increase in stock price, and *ΔEQUITY* is defined as the dollar change in the equity portfolios of CEOs/CFOs with a 1% increase in stock price. It is pertinent to note that Coles et al. (2006) calculated overall managerial incentives without disintegrating the role of CEOs and CFOs. However, motivated by Jiang et al. (2010) and Kim et al. (2011b), we measure CEOs' and CFOs' managerial incentives individually.

[*Insert Table 4.7 here*]

Table 4.7 presents the regression results between investors' sentiment oscillations on StockTwits and managerial incentives. The dependent variable is $NCSKEW_t$. Column (1) presents the results of the interaction between $SENT.OSC_{t-1}$ and incentives of CFO option holdings ($\Delta OPTIONS\_CFO_{\ t-1}$). First, the coefficient of $SENT.OSC_{t-1}$ is positive and statistically significant at 1%, which is consistent with our previous findings. Second, the coefficient of interaction between $SENT.OSC_{t-1}$ and $\Delta OPTIONS\_CFO_{\ t-1}$ is positive and statistically significant at 1%, suggesting that CFO option holdings lure investors on StockTwits to update their economic models (increase in sentiment oscillations), thus augmenting the power of sentiment oscillations to predict future crash risk. Column (2) presents the interaction results between $SENT.OSC_{t-1}$ and the incentives of CEO option holdings ($\Delta OPTIONS\_CEO_{\ t-1}$). The coefficient of interaction is not significant. Similarly, in columns (3) and (4), we examine the interaction of $\Delta EQUITY\_CFO_{\ t-1}$ and $\Delta EQUITY\_CEO_{\ t-1}$ with sentiment oscillations, respectively. The coefficient of interaction is not significant for either CEOs or CFOs. These findings are consistent with Kim et al. (2011b), who present evidence that CFO option incentives increase the crash risk. This is because the options portfolio provides better incentives for managers to increase stock prices artificially. In conjunction with prior literature, we find evidence that CFO options holdings exacerbate the ability of sentiment oscillations to predict future crash risk.

To further test the validity of our findings, we filter the text of the ideas on StockTwits discussing CFOs, CEOs, Options, and Equities (*see Appendix 4.3*). It is interesting to note that investors on StockTwits keenly observe CEOs' and CFOs' trading activities and discuss these with other investors on such platforms, consequently increasing sentiment oscillations. Following Coles et al. (2006), we also measure incentives for all executives who have been offered stock- and equity-based compensation. The results are not reported for brevity reasons. However, the interaction coefficient remains insignificant, further suggesting that CFOs' role is highly influential in bad news hoarding.

4.6.6.2.Financial Reporting Conservatism

Investors on StockTwits interact with each other and discuss firms' financial performance and outlook. For this purpose, firms' financial statements play a pivotal role in ensuring access to all firm-specific information. Therefore, our next avenue to explore triggers of sentiment oscillations on StockTwits is firm-level accounting conservatism. According to Basu (1997), conservatism is accountants' tendency to require stronger verification to recognize good news rather than bad news in firms' financial statements. Watts (2003) recognizes conservatism as a governance mechanism that prevents managers from overstating accounting information provided to all the stakeholders of the firm. In another study, LaFond and Watts (2008) argue that conservatism outweighs the managerial propensity to camouflage bad news and spurs the release of good news in financial markets.

Although we control for firm-level opacity in all our regressions, motivated by these studies and based on the level of extensive discussions on StockTwits about the firm's fair value, we are interested in examining the power of sentiment oscillations to predict crash risk for firms with high/low conservatism. We measure *CSCORE* as a proxy for firm-level conservatism following Khan and Watts (2009) model who extend Basu (1997) time-varying model as follows:

$$X_k = \beta_1 + \beta_2 D_k + \beta_3 R_k + \beta_4 D_k R_k + \varepsilon_k \tag{15}$$

Where *k* represents the firm, *X* is firm-level earnings, R is returns (a proxy for good news), D is a dummy variable that equals one if $R < 0$ and zero otherwise. β_3 measures good news, β_4 is a measure of bad news over good news (conservatism), and total bad news is measured by $\beta_3 + \beta_4$. Khan and Watts (2009) good news and bad news timeliness measures are expressed as a linear function of firm-level characteristics each year, where *GSCORE* represents β_3 and *CSCORE* represents β_4. Both *GSCORE* and *CSCORE* are represented as follows:

$$GSCORE = \gamma_1 + \gamma_2 SIZE + \gamma_3 MTB + \gamma_4 LEV \tag{A}$$

$$CSCORE = \delta_1 + \delta_2 SIZE + \delta_3 MTB + \delta_4 LEV \tag{B}$$

Where $SIZE$ is the market value of the firm's equity, MTB is the market-to-book value of equity, and LEV represents the firm's debt level. Equations (*A*) and (*B*) are not regressions, and instead, they are substituted in Equation (15) as β_3 and β_4, respectively. Finally, *CSCORE* is calculated in annual cross-sectional regressions as follows:

$$\begin{aligned} X_k = \beta_1 + \beta_2 D_k + R_k(\gamma_1 + \gamma_2 SIZE + \gamma_3 MTB + \gamma_4 LEV) \\ + D_k R_k (\delta_1 + \delta_2 SIZE + \delta_3 MTB + \delta_4 LEV) + (\mu_1 SIZE + \mu_2 MTB \\ + \mu_3 LEV + \mu_4 D_k SIZE + \mu_5 D_k MTB + \mu_6 D_k LEV) + \varepsilon_k \end{aligned} \tag{16}$$

Thus, the higher the *CSCORE,* the greater the conservatism. To make the results more intuitive, we divide *CSCORE* into low, medium, and high conservative scores. For each tercile, we run an OLS regression using fixed effects and $NCSKEW_t$ as the dependent variable. We also control for firm-level media coverage and opacity in all regressions, among other control variables. The regression results are presented in Table 4.8, Panel A. The results show that firms with a lower *CSCORE* have a large coefficient of sentiment oscillations, which is statistically significant at 1% compared to firms with a higher *CSCORE,* whose coefficient of sentiment oscillations is not statistically significant. Based on these results, it is evident that investors on StockTwits pay more attention to and discuss those firms with less conservative (lower *CSCORE*) financial statements. This is because, for such firms, the information asymmetry between insiders and outsiders is high, financial statements are opaque and there is a higher tendency of managers to withhold bad news. In contrast, investors actively engage themselves to collect more information about such firms and update their economic models. Therefore, an increase in sentiment oscillations predicts future crash risk for firms with less conservatism. Our findings are consistent with Ball et al. (2012); LaFond and Watts (2008), and Kim and Zhang (2016), who highlight the importance of conservative accounting since it discourages managers from withholding bad news.

[*Insert Table 4.8 here*]

4.6.6.3.Analyst Coverage

Greater analyst coverage plays a pivotal role in reducing information asymmetry between managers and external stakeholders. Therefore, analysts act as monitors and reduce the managerial tendency to manipulate earnings (Yu, 2008). A recent study by Kim et al. (2019), provides evidence that a reduction in firm-level analyst coverage offers significant incentives for managers to withhold bad news. Besides having a monitory role, greater analyst coverage facilitates investors to make investment decisions by providing more industry- and firm-level information (Dyck et al., 2010). In our research setting, we are keen to observe the impact of less analyst coverage on sentiment oscillations and its power to predict future crash risk.

To disentangle the impact of lower and higher analyst coverage, we divide analyst coverage into low, medium, and high terciles. For each tercile, we run an OLS fixed effect regression, where the dependent variable is $NCSKEW_t$. The results in Table 4.8, Panel A, show that the coefficient of sentiment oscillations is statistically significant for firms with less analyst coverage. However, the same is insignificant for firms with greater analyst coverage. These results provide evidence that the impact of sentiment oscillations is more pronounced for firms with less analyst coverage.

4.6.6.4.Competition

Our next avenue in the quest to examine triggers of sentiment oscillations on StockTwits is to investigate the role of product market competition. Giroud and Mueller (2010, 2011) provide evidence that increased product market competition forces managers to become more disciplined. Therefore, product market competition is a vital component of governance mechanisms to discourage self-serving managerial behaviors. Similarly, Callen and Fang (2015b) find a positive association between the short-interest ratio and future crash risk. They suggest that the impact of the short-interest ratio is more pronounced for firms in a less competitive environment.

In our research setting, we are keen to observe the association between sentiment oscillations and future crash risk for firms in a high/low competitive environment. We measure firms' product market competition using the Herfindahl-Hirschman Index (HHI). It is pertinent to note that to make the results more intuitive in Equation (17), we subtract the value of HHI from 1 as follows:

$$COMPETITION = 1 - \sum_{k=1}^{N_j} S_{kjt}^2 \tag{17}$$

Where S_{kjt} is the market share of firm k in industry j in year t. We use firm-level sales as a proxy for market share and use two-digit SIC codes to classify firms into industries. Table 4.8, Panel A presents firms' regression results in low, media, and highly competitive environments. These results show that the impact of sentiment oscillations is more pronounced for firms in a less competitive environment. In conjunction with prior studies, these findings provide further evidence that investors on StockTwits have more sentiment oscillations for firms in a less competitive environment. This is because such firms have weak external monitoring (Callen & Fang, 2015b).

4.6.6.5.Aggregate Market Sentiment

Aggregate market sentiment can influence investors' decision-making in two ways. First, depending on market sentiment, investors become overconfident to predict the accuracy of the signals they receive (Daniel et al., 1998). For example, during high sentiment periods, investors overestimate market signals' accuracy, consequently resulting in market overreactions. Second, such under- and over-reaction because of extreme market sentiment causes excess co-movements among asset prices (Chue et al., 2019). Overall, prior literature provides corroborative evidence suggesting that aggregate market sentiment influences investors' trading behaviors in financial markets (Antoniou et al., 2013, 2016; Stambaugh et al., 2012). In our context, we are interested in examining the moderating impact of aggregate market sentiment to predict future crash risk.

As a proxy for aggregate market sentiment, we use Baker and Wurgler (2006, 2007) sentiment index. Baker-Wurgler use principle component analysis to measure market sentiment based on closed-end fund discount, market turnover (NYSE), first-day returns on IPOs, percentage of equity issues in the issuance of debt and equity, and dividend premium. In addition, they also present an orthogonalized sentiment index based on several macroeconomic indicators. Overall, the BW sentiment index is recognized as a robust proxy for aggregate market sentiment. In a recent study by Huang et al. (2015), they present an alternative proxy for market sentiment measured using partial least squared (PLS) regression (*HJTZ* PLS Sentiment). They argue that the PLS-based sentiment index is a robust measure of market sentiment compared to the BW sentiment index. In Table 4.8, Panel B, we use BW sentiment index, Orthogonalised BW sentiment index, and HJTZ PLS sentiment index. To make the results more intuitive, we divide aggregate market sentiment into negative, neutral, and positive terciles.

Table 4.8, Panel B presents OLS regression results for each tercile. The dependent variable is $NCSKEW_t$. The results show that the coefficient of sentiment oscillations is statistically significant at 1% during positive market sentiment, suggesting that the impact of sentiment oscillations is more able to predict future crash risk during positive sentiment periods. In contrast, we find weak evidence for the impact of negative market sentiment on sentiment oscillations to predict future crash risk. Our findings are consistent with Daniel et al. (1998), who present evidence that investors overestimate their signals' accuracy during high sentiment periods and become overconfident; and Barberis (2018), who suggests that investors' overconfidence convinces investors to update their economic models based on available information. In conjunction with prior studies, these results provide compelling evidence as follows. First, positive market sentiment periods attract investors to participate in the discussions on investor-oriented social media platforms. Second, during extreme market sentiment periods when co-movements among asset prices are high (Chue et al., 2019), investors search for firm-specific information to spread their risk. As a result, they update their sentiment based on available firm-specific information.

4.7. Robustness Checks

4.7.1. Time-span Predictability of Crash Risk

To further check the validity of our results, we use lagged sentiment oscillations to predict future crash risk. This is particularly important for the following reasons. First, according to Roychowdhury and Sletten (2012), bad news hoarding increases across quarters and eventually reaches a peak in the third fiscal quarter. This is because nearing the year-end period initiates scrutiny by stakeholders. Therefore, in our research setting, we expect to notice the same pattern, i.e., an increase in the predictive power of sentiment oscillations from *t-6* to *t-1*. Second, although we test for the impact of unexpected events on crash risk to validate our results in Table 4.4, it is inevitable to test for the implications of unexpected (irrelevant) events that may influence sentiment oscillations in the short term. Therefore, these arguments warrant further evidence to test the robustness of our results.

[*Insert Table 4.9 here*]

We estimate our regression model in Equation (14), and instead of using a one-month lag, we use up to *t-6* lags. Table 4.9 presents the regression results using lagged values of sentiment oscillations. Panel A uses $NCSKEW_t$ as the dependent variable. Columns (1)–(6) represent lagged values of sentiment oscillations from *t-1* to *t-6*, respectively. These results show that the coefficient of sentiment oscillations remains statistically significant at 1% from *t-1* to *t-6*, and the power of sentiment oscillations to predict future crash risk increases as it nears the crash month. It is pertinent to note that these findings suggest that sentiment oscillations can also predict future crash risk as far as six months prior to a crash month, suggesting our results' robustness. Our results in Table 4.9 remain consistent when we use $DUVOL_t$ instead of the negative coefficient of skewness in Panel B.

4.7.2. Alternative measures of Sentiment Oscillations

In our baseline model, we use Maximum Entropy (MaxEnt) to predict investors' sentiment on StockTwits to measure sentiment oscillations. However, there are several machine learning prediction models which can also predict this sentiment. It is pertinent to note that the choice of machine learning approaches depends on the variation in the data set, the size of the training data set, and the expected outcomes of these models. In this regard, to test the validity of our results, we use Random Forest Decisions Trees (RFDT) and Support Vector Machines (SVM) to predict investors' sentiment on StockTwits. We use the same training data set of MaxEnt with the RFDT and SVM models. The accuracy and F1 score of SVM are 80% and 87%, respectively.

Similarly, the accuracy and F1 score of RFDT are 82% and 89%, respectively. We then measure sentiment oscillations using a similar method to that in Equation (8). Table 4.10, Panel A presents our regression results after using alternative machine learning approaches to measure sentiment oscillations. Our results remain consistent after using these alternative models.

[*Insert Table 4.10 here*]

4.7.3. Alternative Crash Risk measures

There are several methods to measure crash risk. To check the robustness of our results, we include three additional measures of crash risk. Our first measure is the negative coefficient of minimum return ($NCMRET_{k,t}$), which is a continuous variable, and the other two measures are *CRASH320* and *CRASH20,* which are dichotomous variables. The calculation methods for these measures are explained in section 4. We use OLS regression with fixed effects for continuous dependent variables; and for dichotomous dependent variables, we use logit regression with fixed effects. Table 4.10, Panel B presents the regression results. For alternative crash risk measures, the coefficient of sentiment oscillations is statistically significant at 1%. These results further suggest that our findings are robust to using alternative crash risk measures.

4.8. Chapter Conclusion

Using a large sample of US public firms from 2012–2017 and using more than 13 million ideas posted on StockTwits while discussing the sample firms, we investigate the association between sentiment oscillations on StockTwits and firm-level future crash risk. To measure sentiment oscillations, we use the Maximum Entropy Method to predict investors' sentiment and use the log of a 3-month moving average of the sum of the number of times a distinct investor updates her sentiment aggregated at the firm-month level. Each update in investors' sentiment is assigned a weight based on the number of followers each investor has on their StockTwits' profile. This study provides corroborative evidence that an increase in sentiment oscillations on StockTwits predicts firm-level future crash risk. These results hold after mitigating the impact of unexpected market events. Moreover, these results present that professional investors' sentiment oscillations have more predictive power than novice investors' sentiment oscillations. Our results remain robust when controlling for firm, industry, and time fixed effects and addressing endogeneity concerns using the 2SLS instrumental variable approach.

We conduct several tests to investigate the role of various channels which trigger sentiment oscillations on StockTwits. We find that the impact of sentiment oscillations is more pronounced when there is a greater sensitivity of the CFO options portfolio with a 1% change in stock price, less conservative financial reporting by managers, less analyst coverage, and a low level of product market competition. Moreover, positive sentiment in the market attracts investors to participate in the discussion, consequently increasing the number of sentiment oscillations to predict firm-level future crash risk.

To the best of our knowledge, this is the first study that examines the relationship between firm-level crash risk and investor-oriented social media platforms. Our study contributes to the existing literature on crash risk: first, by offering a unique proxy for sentiment oscillations predict future crash risk; second, by providing firsthand evidence for the heterogeneity of investors in financial markets; third, by highlighting the significance of investor-oriented social media platforms to influence investors' opinions in the financial markets.

This study has practical implications for investors, regulators, and professional money managers. For investors, StockTwits offers a unique opportunity to sift through an abundance of information, follow influential investors and analysts, and learn the tricks and techniques of investing in financial markets. Therefore, investors can gain better market insights and, overall, the cost of information acquisition on such platforms is minimal. Investor-oriented social media platforms pose a challenge for the regulators since there is no regulatory monitoring mechanism for such platforms. This study highlights the significant role of social media platforms. If well-regulated, such platforms can play a critical role in enhancing the efficiency of financial markets. By leveraging social media platforms, professional money managers have developed sentiment indices to gauge investors' sentiment in financial markets. However, our study highlights another key benefit of consuming data from social media platforms for investors, i.e., predicting firm-wide events to adjust their portfolio betas and improve their analyses.

Table 4.1: Summary Statistics

Notes: The table presents the summary statistics of all variables of interest. Panel A presents summary statistics of six crash risk measures. ***$NCSKEW_t$*** is the negative coefficient of the skewness, ***$DUVOL_t$*** is the degree of up and down volatility of the firm-level returns, ***$NCMRET_t$*** is the negative coefficient of the minimum return, ***$CRASH309_t$*** is the dummy variable equals one if the weekly return decline more than 3.09 standard deviation within the given month, ***$CRASH320_t$*** is the dummy variable equals one if the weekly return decline more than 3.20 standard deviation within the given month and, ***$CRASH20_t$*** is the dummy variable equals one if the firm experience at least one market-adjusted weekly return less than 20% within the given month. Panel B presents firm-level sentiment oscillations (***$SENT.OSC_{t-1}$***), sentiment oscillations for professional (***$SENT.OSC\_Pro_{t-1}$***) and novice (***$SENT.OSC\_Novice_{t-1}$***) investors, and sentiment oscillation at industry levels (***$SENT\_OSC\_Ind_{t-1}$***). Whereas the investors' sentiment is based on the Maximum Entropy prediction model (MaxEnt). After calculating the investors' sentiment, the sentiment oscillations are measured as the log of the 3-months moving average of the sum of the number of times a distinct investor updates her sentiment (bullish-to-bearish or vice versa) aggregated at the firm-month level. Panel C and D present the summary statistics of baseline controls and additional variables. The definitions of these variables are presented in Appendix 4.1.

	Mean	SD	P10	P25	Median	P75	Max
Panel A: Crash Risk							
$NCSKEW_t$	0.075	1.981	-2.525	-1.864	0.152	1.983	3.381
$DUVOL_t$	0.089	2.123	-2.668	-1.499	0.099	1.668	6.346
$NCMRET_t$	2.320	0.850	1.365	1.714	2.178	2.793	7.114
$CRASH309_t$	0.056	0.230	0.000	0.000	0.000	0.000	1.000
$CRASH320_t$	0.045	0.208	0.000	0.000	0.000	0.000	1.000
$CRASH20_t$	0.018	0.134	0.000	0.000	0.000	0.000	1.000
Panel B: Sentiment Oscillations							
$SENT.OSC_{t-1}$	1.762	1.336	0.288	0.847	1.466	2.428	6.986
$SENT.OSC\_Pro_{t-1}$	1.078	0.961	0.000	0.405	0.847	1.540	4.790
$SENT.OSC\_Novice_{t-1}$	1.173	1.144	0.000	0.288	0.847	1.674	5.752
$SENT.OSC\_Ind_{t-1}$	12.652	3.746	7.323	10.041	12.926	16.123	19.047
Panel C: Baseline Controls							
$NCSKEW_{t-1}$	0.071	1.980	-2.525	-1.867	0.144	1.978	3.381
$SIGMA_{t-1}$	0.046	0.034	0.015	0.023	0.037	0.058	0.321
RET_{t-1}	0.003	0.026	-0.026	-0.010	0.003	0.017	0.139
$DTURN_{t-1}$	0.002	0.056	-0.045	-0.017	0.000	0.018	0.242
$KURT_{t-1}$	1.917	0.435	1.318	1.606	1.955	2.181	3.207
$MEDIA_{t-1}$	1.461	0.792	0.000	0.693	1.609	2.079	3.738
ROA_{t-1}	-0.001	0.051	-0.051	-0.004	0.010	0.022	0.097
BTM_{t-1}	0.416	0.358	0.076	0.183	0.330	0.546	2.111
LEV_{t-1}	0.528	0.259	0.195	0.346	0.515	0.680	1.417
$OPACITY_{t-1}$	0.088	0.156	0.000	0.000	0.000	0.114	1.055
$GOOD_{t-1}$	0.149	0.154	0.000	0.007	0.106	0.246	0.594
$SIZE_{t-1}$	5.602	2.041	3.119	4.481	5.731	6.917	10.173
RD_{t-1}	0.015	0.029	0.000	0.000	0.000	0.019	0.172
AGE10	0.224	0.417	0.000	0.000	0.000	0.000	1.000
Panel D: Additional Variables							
$\Delta OPTIONS\_CEO_{t-1}$	0.281	0.722	0.000	0.001	0.071	0.260	18.672
$\Delta OPTIONS\_CFO_{t-1}$	0.068	0.387	0.000	0.000	0.016	0.058	13.479
$\Delta EQUITY\_CEO_{t-1}$	1.047	11.986	0.012	0.040	0.105	0.301	467.949
$\Delta EQUITY\_CFO_{t-1}$	0.041	0.097	0.001	0.006	0.017	0.043	2.547
$CSCORE_t$	0.179	0.207	-0.050	0.059	0.166	0.279	1.534
$ANALYSTS_t$	1.964	0.812	0.693	1.386	2.079	2.565	3.434
$COMPETITION_t$	0.215	0.025	0.200	0.201	0.203	0.218	0.376
$BW.SENT_t$	-0.204	0.147	-0.372	-0.315	-0.225	-0.118	0.207
$BW.SENT\_Orth_t$	-0.055	0.118	-0.207	-0.150	-0.075	0.034	0.181
$PLS.SENT_t$	-0.715	0.137	-0.916	-0.806	-0.687	-0.609	-0.470

Table 4.2: Pearson Correlation Coefficients

Notes: The table presents the Pearson correlation coefficients of baseline variables in Panel A, and Panel B presents the Pearson correlation coefficients among the six crash risk measures. *, **, *** represent significance at 10%, 5% and 1% level, respectively.

Panel A: Correlation coefficients between baseline variables

	$NCSKEW_t$	$SENT.OSC_{t-1}$	$NCSKEW_{t-1}$	$SIGMA_{t-1}$	RET_{t-1}	$DTURN_{t-1}$	$KURT_{t-1}$	$MEDIA_{t-1}$	ROA_{t-1}	BTM_{t-1}	LEV_{t-1}	$OPACITY_{t-1}$	$GOOD_{t-1}$	$SIZE_{t-1}$	RD_{t-1}
$SENT.OSC_{t-1}$	0.021***														
$NCSKEW_{t-1}$	-0.015***	0.021***													
$SIGMA_{t-1}$	0.009***	0.186***	-0.045***												
RET_{t-1}	0.065***	-0.019***	-0.660***	0.083***											
$DTURN_{t-1}$	0.007**	0.129***	0.005	0.153***	0.006*										
$KURT_{t-1}$	0.014***	0.018***	-0.024***	0.134***	0.013***	0.011***									
$MEDIA_{t-1}$	-0.008**	0.145***	-0.014***	0.020***	0.030***	0.009***	0.022***								
ROA_{t-1}	-0.001	-0.159***	-0.009***	-0.346***	0.018***	-0.054***	-0.010***	0.110***							
BTM_{t-1}	-0.068***	-0.119***	0.012***	0.109***	-0.092***	0.007**	0.006*	-0.066***	-0.001						
LEV_{t-1}	-0.000	0.038***	0.001	-0.041***	0.001	0.017***	-0.008**	0.076***	0.007**	-0.248***					
$OPACITY_{t-1}$	0.012***	0.221***	0.012***	0.179***	-0.029***	-0.029***	0.009***	-0.049***	-0.266***	-0.048***	0.004				
$GOOD_{t-1}$	-0.009***	-0.114***	-0.009***	-0.196***	-0.002	-0.031***	-0.001	0.066***	0.171***	-0.004	0.087***	-0.102***			
$SIZE_{t-1}$	-0.013***	0.098***	-0.012***	-0.381***	-0.007**	-0.061***	-0.022***	0.283***	0.522***	0.010***	0.366***	-0.260***	0.239***		
RD_{t-1}	0.003	0.191***	0.001	0.269***	0.021***	0.042***	0.012***	-0.070***	-0.639***	-0.190***	-0.184***	0.263***	-0.195***	-0.554***	
$AGE10$	0.009***	0.048***	0.010***	0.191***	-0.003	0.023***	0.003	-0.100***	-0.231***	-0.016***	-0.148***	0.163***	-0.074***	-0.341***	0.238***

Panel B: Correlation coefficients between crash risk variables

	$NCSKEW_t$	$DUVOL_t$	$NCMRET_t$	$CRASH309_t$	$CRASH320_t$	$CRASH20_t$
$NCSKEW_t$	1					
$DUVOL_t$	0.946***	1				
$NCMRET_t$	0.029***	0.033***	1			
$CRASH309_t$	0.154***	0.152***	0.149***	1		
$CRASH320_t$	0.141***	0.140***	0.147***	0.873***	1	
$CRASH20_t$	0.101***	0.106***	0.143***	0.124***	0.123***	1

Table 4.3: Sentiment Oscillations and Crash Risk

Notes: The table presents the regression results of sentiment oscillations and crash risk based on Equation (14). Columns 1, 2, and 3, 4 present results from ordinary least square regression using fixed effects where the dependent variables are ***NCSKEW$_t$*** defined as the negative coefficient of the skewness and ***DUVOL$_t$*** defined as the degree of up and down the volatility of the firm-level returns, respectively. Columns 5 and 6 present the results from logistic regression where the dependent variable is ***CRASH309$_t$***, which is defined as the dummy variable that equals one if the weekly return decline more than 3.09 standard deviation within the given month. The independent variable in columns (1) – (6) is the sentiment oscillations ***(SENT.OSC$_{t-1}$)*** measured as the log of the 3-months moving average of the sum of the number of times a distinct investor updates her sentiment (bullish-to-bearish or vice versa) aggregated at the firm-month level. The definitions of all variables are presented in Appendix 4.1. Columns 1, 3, and 5 use firm and time fixed effects, and columns 2, 4 and, 6 use industry and time fixed effects, respectively. All continuous variables are winsorized at 1st and 99th percentiles and standardized. *, **, *** represent significance at 10%, 5% and 1% level, respectively. The standard errors are reported in brackets and are robust to heteroscedasticity and clustered at the firm-level.

	NCSKEW$_t$		***DUVOL$_t$***		***CRASH309$_t$***	
	1	**2**	**3**	**4**	**5**	**6**
	0.039***	0.040***	0.036***	0.036***	0.057***	0.062***
SENT.OSC$_{t-1}$	[0.006]	[0.006]	[0.007]	[0.007]	[0.016]	[0.017]
	-0.026**	-0.027**	-0.007	-0.008	0.018	0.060
SENT.OSC_Ind$_{t-1}$	[0.011]	[0.011]	[0.013]	[0.013]	[0.016]	[0.042]
	0.030***	0.032***	0.029***	0.032***	0.087***	0.090***
NCSKEW$_{t-1}$	[0.004]	[0.004]	[0.005]	[0.005]	[0.018]	[0.018]
	0.003	0.003	-0.003	-0.003	0.027	0.028
SIGMA$_{t-1}$	[0.004]	[0.004]	[0.005]	[0.005]	[0.017]	[0.017]
	0.073***	0.073***	0.073***	0.073***	-0.046**	-0.042**
RET$_{t-1}$	[0.004]	[0.004]	[0.005]	[0.005]	[0.019]	[0.019]
	0.014***	0.014***	0.020***	0.019***	0.068***	0.066***
KURT$_{t-1}$	[0.003]	[0.003]	[0.004]	[0.004]	[0.014]	[0.014]
	0.004	0.005	-0.003	-0.002	-0.030**	-0.029**
DTURN$_{t-1}$	[0.004]	[0.004]	[0.004]	[0.004]	[0.014]	[0.014]
	0.006	0.006	0.007	0.008*	-0.316***	-0.315***
MEDIA$_{t-1}$	[0.004]	[0.004]	[0.005]	[0.005]	[0.015]	[0.015]
	0.006	0.006	-0.003	-0.004	0.077***	0.047**
ROA$_{t-1}$	[0.006]	[0.006]	[0.007]	[0.007]	[0.021]	[0.021]
	-0.263***	-0.260***	-0.254***	-0.251***	-0.144***	-0.127***
BTM$_{t-1}$	[0.012]	[0.012]	[0.013]	[0.013]	[0.017]	[0.018]
	-0.094***	-0.094***	-0.084***	-0.083***	-0.091***	-0.069***
LEV$_{t-1}$	[0.010]	[0.010]	[0.012]	[0.012]	[0.017]	[0.017]
	0.011***	0.011***	0.013**	0.012**	0.045***	0.042***
OPACITY$_{t-1}$	[0.004]	[0.004]	[0.005]	[0.005]	[0.015]	[0.015]
	0.028**	0.030**	0.016	0.020	0.063***	0.025
GOOD$_{t-1}$	[0.012]	[0.012]	[0.014]	[0.014]	[0.014]	[0.016]
	0.045**	0.045**	0.032	0.029	0.146***	0.155***
SIZE$_{t-1}$	[0.022]	[0.022]	[0.025]	[0.025]	[0.022]	[0.024]
	-0.025**	-0.027***	-0.034***	-0.037***	0.049**	0.029
RD$_{t-1}$	[0.010]	[0.010]	[0.012]	[0.012]	[0.021]	[0.022]
	-0.018	-0.020	-0.007	-0.010	-0.088**	-0.058
AGE10	[0.019]	[0.019]	[0.024]	[0.024]	[0.036]	[0.041]
Industry FE	N	Y	N	Y	N	Y
Firm FE	Y	N	Y	N	Y	N
Time FE	Y	Y	Y	Y	Y	Y
R-squared	0.037	0.052	0.046	0.064		
Pseudo R-squared					0.032	0.037
Firms	1815	1815	1814	1814	1815	1815
Observations	104,137	104,137	72,315	72,315	103,536	103,536

Table 4.4: Sentiment Oscillations on Crash Risk during Earnings Announcements

Notes: The table presents the regression results of the sentiment oscillations and crash risk during quarterly earnings announcements. Columns *QEA* contains the sample of firms where the crash risk month coincides with the quarterly earnings announcement month, and *Non-QEA* is the sample of firms where the crash risk month does not coincide with the quarterly earnings announcements. The dependent variable $\boldsymbol{NCSKEW_t}$ is the negative coefficient of the skewness, $\boldsymbol{DUVOL_t}$ is the degree of up and down volatility of the firm-level returns, and $\boldsymbol{CRASH309_t}$ is the dummy variable that equals one if the weekly return decline more than 3.09 standard deviation within the given month, respectively. In the models where dependent variables are continuous ($\boldsymbol{NCSKEW_t}$ and $\boldsymbol{DUVOL_t}$), we use ordinary least square regression using industry and time fixed effects. In the model where the dependent variable is dichotomous ($\boldsymbol{CRASH309_t}$), we use logistics regression using industry and time fixed effects. The independent variable is the sentiment oscillations ($\boldsymbol{SENT.OSC_{t-1}}$), measured as the log of the 3-months moving average of the sum of the number of times a distinct investor updates her sentiment (bullish-to-bearish or vice versa) aggregated at the firm-month level. Quarterly earnings announcement dates are collected from Compustat. The definitions of all variables are presented in Appendix 4.1. All continuous variables are winsorized at 1st and 99th percentiles and standardized. *, **, *** represent significance at 10%, 5% and 1% level, respectively. The standard errors are reported in brackets and are robust to heteroscedasticity and clustered at the firm-level.

	$NCSKEW_t$		$DUVOL_t$		$CRASH309_t$	
	1	*2*	*3*	*4*	*5*	*6*
	QEA	*Non-QEA*	*QEA*	*Non-QEA*	*QEA*	*Non-QEA*
$SENT.OSC_{t-1}$	0.065***	0.027***	0.057***	0.025***	0.041*	0.050*
	[0.011]	[0.007]	[0.014]	[0.009]	[0.022]	[0.028]
$SENT.OSC\_Ind_{t-1}$	0.013	-0.043***	0.043*	-0.027*	0.124**	0.009
	[0.020]	[0.013]	[0.025]	[0.016]	[0.054]	[0.069]
$NCSKEW_{t-1}$	0.060***	0.031***	0.061***	0.030***	0.084***	0.110***
	[0.008]	[0.005]	[0.009]	[0.006]	[0.023]	[0.031]
$SIGMA_{t-1}$	-0.011	0.011**	-0.022**	0.006	0.054**	0.062**
	[0.008]	[0.005]	[0.010]	[0.006]	[0.024]	[0.025]
RET_{t-1}	0.082***	0.068***	0.078***	0.070***	-0.100***	0.040
	[0.008]	[0.005]	[0.010]	[0.006]	[0.026]	[0.030]
$KURT_{t-1}$	0.008	0.014***	0.016**	0.020***	0.060***	0.090***
	[0.006]	[0.004]	[0.007]	[0.005]	[0.017]	[0.023]
$DTURN_{t-1}$	-0.002	0.008*	-0.003	-0.001	-0.054***	0.020
	[0.006]	[0.004]	[0.008]	[0.005]	[0.018]	[0.022]
$MEDIA_{t-1}$	0.000	0.004	-0.009	0.009	-0.092***	-0.113***
	[0.008]	[0.005]	[0.010]	[0.006]	[0.022]	[0.026]
ROA_{t-1}	-0.019*	0.018**	-0.034***	0.009	0.083***	0.039
	[0.011]	[0.007]	[0.013]	[0.009]	[0.029]	[0.033]
BTM_{t-1}	-0.298***	-0.241***	-0.288***	-0.235***	-0.113***	-0.110***
	[0.015]	[0.013]	[0.018]	[0.014]	[0.023]	[0.030]
LEV_{t-1}	-0.154***	-0.068***	-0.148***	-0.056***	-0.078***	-0.027
	[0.016]	[0.011]	[0.020]	[0.013]	[0.023]	[0.028]
$OPACITY_{t-1}$	-0.022***	0.025***	-0.020**	0.025***	0.047**	0.034
	[0.008]	[0.005]	[0.010]	[0.006]	[0.020]	[0.023]
$GOOD_{t-1}$	0.018	0.033**	0.008	0.023	0.034*	0.004
	[0.023]	[0.015]	[0.027]	[0.017]	[0.020]	[0.028]
$SIZE_{t-1}$	0.096***	0.025	0.071	0.026	0.130***	0.009
	[0.035]	[0.025]	[0.045]	[0.030]	[0.031]	[0.039]
RD_{t-1}	-0.044**	-0.019	-0.060***	-0.024*	-0.066**	0.092***
	[0.017]	[0.012]	[0.023]	[0.015]	[0.031]	[0.033]
AGE10	-0.035	-0.014	-0.031	0.004	-0.026	-0.013
	[0.033]	[0.021]	[0.041]	[0.027]	[0.053]	[0.068]
Fixed Effects	Y	Y	Y	Y	Y	Y
R-squared	0.095	0.066	0.128	0.079		
Pseudo R-squared					0.031	0.015
Firms	1809	1815	1806	1813	1809	1815
Observations	34,244	69,772	22,474	49,732	33,986	69,279

Table 4.5: Investors' Experience on StockTwits and Crash Risk

Notes: The table presents the regression results of sentiment oscillation based on investors' experience on StockTwits and crash risk. Based on voluntarily disclosed investors' experience, we classify investors into two groups. First, Professional investors are those who are experienced investment professionals. Second, Novice investors are those who classify themselves as either intermediate or novice. The dependent variable ***NCSKEW$_t$*** is the negative coefficient of the skewness, ***DUVOL$_t$*** is the degree of up and down volatility of the firm-level returns, and ***CRASH309$_t$*** is the dummy variable that equals one if the weekly return decline more than 3.09 standard deviation within the given month, respectively. The models where dependent variables are continuous (NCSKEWt and DUVOLt) use ordinary least square regression using industry and time fixed effects. In the model where the dependent variable is dichotomous (***CRASH309$_t$***), we use logistics regression using industry and time fixed effects. The independent variable is the sentiment oscillations (***SENT.OSC$_{t-1}$***), measured as the log of the 3-months moving average of the sum of the number of times a distinct investor updates her sentiment (bullish-to-bearish or vice versa) aggregated at the firm-month level. The definitions of all variables are presented in Appendix 4.1. All continuous variables are winsorized at 1st and 99th percentiles and standardized. *, **, *** represent significance at 10%, 5% and 1% level, respectively. The standard errors are reported in brackets and are robust to heteroscedasticity and clustered at the firm-level.

	NCSKEW$_t$			*DUVOL$_t$*			*CRASH309$_t$*		
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
	Professional	*Novice*	*Both*	*Professional*	*Novice*	*Both*	*Professional*	*Novice*	*Both*
SENT.OSC_Pro$_{t-1}$	0.037***		0.031***	0.036***		0.036***	0.062***		0.051*
	[0.006]		[0.007]	[0.007]		[0.009]	[0.017]		[0.028]
SENT.OSC_Novice$_{t-1}$		0.021***	0.005		0.013*	-0.004		0.055***	0.024
		[0.006]	[0.008]		[0.008]	[0.009]		[0.017]	[0.028]
SENT.OSC_Ind$_{t-1}$	-0.014	-0.016	-0.023*	0.001	0.014	0.001	0.104**	0.116**	0.135***
	[0.011]	[0.011]	[0.012]	[0.013]	[0.014]	[0.015]	[0.043]	[0.046]	[0.048]
NCSKEW$_{t-1}$	0.033***	0.034***	0.035***	0.033***	0.036***	0.036***	0.089***	0.104***	0.095***
	[0.004]	[0.005]	[0.005]	[0.005]	[0.006]	[0.006]	[0.019]	[0.020]	[0.021]
SIGMA$_{t-1}$	0.004	0.001	0.001	-0.002	-0.003	-0.003	0.028	0.035*	0.030
	[0.004]	[0.004]	[0.004]	[0.005]	[0.005]	[0.005]	[0.018]	[0.018]	[0.019]
RET$_{t-1}$	0.073***	0.072***	0.072***	0.072***	0.072***	0.070***	-0.039*	-0.033	-0.034
	[0.004]	[0.005]	[0.005]	[0.005]	[0.006]	[0.006]	[0.020]	[0.021]	[0.022]
KURT$_{t-1}$	0.013***	0.013***	0.012***	0.019***	0.019***	0.017***	0.065***	0.065***	0.059***
	[0.003]	[0.003]	[0.004]	[0.004]	[0.004]	[0.004]	[0.014]	[0.015]	[0.016]
DTURN$_{t-1}$	0.005	0.004	0.002	-0.002	-0.002	-0.005	-0.030**	-0.029**	-0.033**
	[0.004]	[0.004]	[0.004]	[0.005]	[0.005]	[0.005]	[0.015]	[0.015]	[0.015]
MEDIA$_{t-1}$	0.007*	0.009**	0.009**	0.011**	0.008	0.010*	-0.318***	-0.333***	-0.338***
	[0.004]	[0.004]	[0.004]	[0.005]	[0.005]	[0.005]	[0.016]	[0.017]	[0.017]
ROA$_{t-1}$	0.005	0.012*	0.013*	-0.005	0.002	0.003	0.053**	0.055**	0.060**
	[0.007]	[0.007]	[0.007]	[0.008]	[0.008]	[0.008]	[0.022]	[0.023]	[0.024]
BTM$_{t-1}$	-0.259***	-0.266***	-0.266***	-0.252***	-0.261***	-0.264***	-0.122***	-0.128***	-0.125***
	[0.013]	[0.013]	[0.013]	[0.013]	[0.014]	[0.014]	[0.019]	[0.020]	[0.022]
LEV$_{t-1}$	-0.095***	-0.097***	-0.099***	-0.088***	-0.092***	-0.097***	-0.062***	-0.060***	-0.056***
	[0.010]	[0.011]	[0.011]	[0.012]	[0.013]	[0.013]	[0.018]	[0.019]	[0.020]
OPACITY$_{t-1}$	0.013***	0.011**	0.012**	0.012**	0.012**	0.012**	0.044***	0.043***	0.040**
	[0.004]	[0.005]	[0.005]	[0.005]	[0.006]	[0.006]	[0.016]	[0.016]	[0.016]
GOOD$_{t-1}$	0.036***	0.038***	0.040***	0.025*	0.023	0.025	0.017	0.018	0.009
	[0.013]	[0.014]	[0.015]	[0.014]	[0.016]	[0.017]	[0.017]	[0.018]	[0.019]
SIZE$_{t-1}$	0.049**	0.039*	0.033	0.040	0.031	0.030	0.142***	0.157***	0.140***
	[0.022]	[0.023]	[0.023]	[0.026]	[0.027]	[0.027]	[0.025]	[0.026]	[0.027]
RD$_{t-1}$	-0.025**	-0.018*	-0.016	-0.036***	-0.023*	-0.023*	0.030	0.036	0.032
	[0.010]	[0.011]	[0.011]	[0.013]	[0.013]	[0.014]	[0.023]	[0.024]	[0.025]
AGE10	-0.023	-0.044**	-0.049**	-0.022	-0.016	-0.027	-0.060	-0.037	-0.049
	[0.020]	[0.020]	[0.021]	[0.025]	[0.026]	[0.027]	[0.044]	[0.046]	[0.048]
Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
R-squared	0.053	0.057	0.058	0.067	0.071	0.074			
Pseudo R-squared							0.038	0.038	0.039
Firms	1814	1813	1813	1813	1812	1811	1814	1813	1813
Observations	96,717	86,119	81,250	67,178	59,808	56,432	96,171	85,609	80,712

Table 4.6: Testing for Endogeneity using 2SLS Regression

Notes: The table presents results from two stages least square regression (2SLS) using the instrumental variable approach. Our first instrument is the psychic distances (***PSYCH_DIST$_{t-1}$***) among investors who are posting ideas about the specific firms in our sample. Since we are using the US sample firms, we use the United States as the base country and then use Dow and Karunaratna's (2006) psychic distance measures using investors' self-disclosed locations at the country-level excluding the United States on StockTwits. The second instrument is ***PROXIMITY$_{t-1}$***, defined as the distance between investors' location at the city-level within the United States and the city of the firms' headquarter. Sanderson-Windmeijer F test of excluded instruments is presented as S-W F-statistics, and Kleibergen-Paap Wald F statistic is presented as K-P Wald F-statistics. In the first stage, the dependent variable is the firm-level sentiment oscillations ***(SENT.OSC$_{t-1}$)***. In the second stage, the dependent variables are ***NCSKEW$_t$***, defined as the negative coefficient of the skewness, and ***DUVOL$_t$*** as the degree of up and down volatility of the firm-level returns. We use industry and time fixed effects in all the regressions. The definitions of all variables are presented in Appendix 4.1. All continuous variables are winsorized at 1st and 99th percentiles and standardized. *, **, *** represent significance at 10%, 5% and 1% level, respectively. The standard errors are reported in brackets and are robust to heteroscedasticity and clustered at the firm-level.

	First Stage	*Second Stage*	
		NCSKEW$_t$	***DUVOL$_t$***
SENT.OSC$_{t-1}$		0.038***	0.040**
		[0.015]	[0.017]
PSYCH_DIST$_{t-1}$	0.091***		
	[0.002]		
PROXIMITY$_{t-1}$	0.171***		
	[0.012]		
NCSKEW$_{t-1}$	0.011***	0.032***	0.033***
	[0.003]	[0.004]	[0.005]
SIGMA$_{t-1}$	0.086***	0.006	-0.002
	[0.004]	[0.004]	[0.005]
RET$_{t-1}$	-0.015***	0.074***	0.074***
	[0.004]	[0.004]	[0.005]
KURT$_{t-1}$	-0.001	0.014***	0.020***
	[0.002]	[0.003]	[0.004]
DTURN$_{t-1}$	-0.021***	-0.009***	-0.010***
	[0.006]	[0.003]	[0.004]
MEDIA$_{t-1}$	-0.017***	0.008**	0.009**
	[0.002]	[0.004]	[0.005]
ROA$_{t-1}$	-0.032***	0.006	-0.003
	[0.009]	[0.006]	[0.007]
BTM$_{t-1}$	-0.050***	-0.259***	-0.251***
	[0.015]	[0.012]	[0.013]
LEV$_{t-1}$	0.107***	-0.095***	-0.084***
	[0.021]	[0.010]	[0.012]
OPACITY$_{t-1}$	0.122***	0.008*	0.011*
	[0.009]	[0.005]	[0.006]
GOOD$_{t-1}$	0.081***	0.028**	0.019
	[0.020]	[0.012]	[0.014]
SIZE$_{t-1}$	0.373***	0.041*	0.026
	[0.044]	[0.022]	[0.026]
RD$_{t-1}$	-0.059***	-0.026**	-0.036***
	[0.021]	[0.010]	[0.012]
Fixed Effects	Y	Y	Y
Firms	1815	1815	1815
Observations	104137	104137	72315
S-W F-statistics	1172.60***	1010.07***	
K-P Wald F-statistics	1172.60***	1010.07***	
Sargan P-Value		0.831	0.388

Table 4.7: Sentiment Oscillations, Managerial Incentives, and Crash Risk

Notes: The table presents the regression results of the moderating effect of managerial incentives on sentiment oscillations and crash risk. We use ordinary least square (OLS) regression using industry and time fixed effects. The dependent variable ***$NCSKEW_t$*** is the negative coefficient of the skewness, and the independent variable is the sentiment oscillations ***($SENT.OSC_{t-1}$)***. Column (1) and (2) presents the regression results based on the interaction between sentiment oscillations (***$SENT.OSC_{t-1}$***) and ***$\Delta OPTIONS\_CFO_{t-1}$***, defined as the dollar change in the value of CFO option holdings resulting from a 1% increase in the firm stock price and sentiment oscillations (***$SENT.OSC_{t-1}$***) and ***$\Delta OPTIONS\_CEO_{t-1}$***, defined as the dollar change in the value of CEO options holdings resulting from a 1% increase in the firm stock price, respectively. Similarly, columns (3) and (4) presents the regression results based on the interaction between sentiment oscillations (***$SENT.OSC_{t-1}$***) and ***$\Delta EQUITY\_CFO_{t-1}$***, defined as the dollar change in the value of CFO equity ownership resulting from a 1% increase in the firm stock price and sentiment oscillations (***$SENT.OSC_{t-1}$***) and ***$\Delta EQUITY\_CEO_{t-1}$***, defined as the dollar change in the value of CEO equity ownership resulting from a 1% increase in the firm stock price, respectively. The definitions of all variables are presented in Appendix 4.1. All continuous variables are winsorized at 1st and 99th percentiles and standardized. *, **, *** represent significance at 10%, 5% and 1% level, respectively. The standard errors are reported in brackets and are robust to heteroscedasticity and clustered at the firm-level.

$NCSKEW_t$	***1***	***2***	***3***	***4***
$SENT.OSC_{t-1}$	0.047*** [0.008]	0.051*** [0.008]	0.047*** [0.008]	0.052*** [0.008]
$\Delta OPTIONS\_CFO_{t-1}$	-0.028*** [0.010]			
*$SENT.OSC_{t-1}$ * $\Delta OPTIONS\_CFO_{t-1}$*	0.011*** [0.003]			
$\Delta OPTIONS\_CEO_{t-1}$		-0.035*** [0.008]		
*$SENT.OSC_{t-1}$ * $\Delta OPTIONS\_CEO_{t-1}$*		0.006 [0.004]		
$\Delta EQUITY\_CFO_{t-1}$			-0.020** [0.010]	
*$SENT.OSC_{t-1}$ * $\Delta EQUITY\_CFO_{t-1}$*			0.003 [0.006]	
$\Delta EQUITY\_CEO_{t-1}$				-0.001 [0.007]
*$SENT.OSC_{t-1}$ * $\Delta EQUITY\_CEO_{t-1}$*				-0.010 [0.007]
$SENT\_OSC\_Ind_{t-1}$	-0.032** [0.013]	-0.037*** [0.013]	-0.031** [0.013]	-0.036*** [0.013]
CONTROLS	Y	Y	Y	Y
Fixed Effects	Y	Y	Y	Y
R-squared	0.052	0.053	0.052	0.053
Firms	1190	1175	1190	1175
Observations	68,293	68,484	68,293	68,484

Table 4.8: External monitoring by Social Media and Crash Risk

Notes: The table presents the regression results of the moderating effect of the various firm- and market-level variables and their impact on sentiment oscillations and crash risk. We use ordinary least square (OLS) regression using industry and time fixed effects in all the models. Panel A presents the moderating effect of firm-level ***Conservatism*** proxied by ***C Score*** calculated by estimating a firm-level cross-sectional regression between earnings and the firm size, market-to-book ratio, and leverage; ***Analyst Coverage*** measured by the number of analysts covering the firm and, ***Competition*** measured by Herfindahl-Hirschman Index (HHI) of firms' quarterly total assets. Each moderator is divided into equal terciles of the low, medium, and high. Panel B presents the regression results from the moderating effect of the market sentiment on sentiment oscillations and crash risk. We use Baker and Wurgler (2006) (***BW***) and Huang et al. (2015) (***HJTZ***) sentiments indices, whereas ***BW Sentiment*** is the simple market sentiment before incorporating the impact of macroeconomic factors, and ***Orthogonalised BW Sentiment*** is the BW sentiment after incorporating the impact of macroeconomic factors. Finally, ***HJTZ PLS sentiment*** is the partial least square sentiment index. HJTZ PLS sentiment data is only available until 2016. Each moderator (market sentiment proxy) is divided into negative, neutral, and positive market sentiment. The dependent variable ***$NCSKEW_t$*** is the negative coefficient of the skewness, and the independent variable is the sentiment oscillations (***$SENT.OSC_{t-1}$***) in all the regressions. The definitions of all variables are presented in Appendix 4.1. All continuous variables are winsorized at 1st and 99th percentiles and standardized. *, **, *** represent significance at 10%, 5% and 1% level, respectively. The standard errors are reported in brackets and are robust to heteroscedasticity and clustered at the firm-level.

Panel A: Moderating effect of the firm-level variables on Sentiment Oscillations and Crash Risk

	Conservatism				*Analyst Coverage*				*Competition*		
$NCSKEW_t$	*LOW*	*MEDIUM*	*HIGH*	***$NCSKEW_t$***	*LOW*	*MEDIUM*	*HIGH*	***$NCSKEW_t$***	*LOW*	*MEDIUM*	*HIGH*
$SENT.OSC_{t-1}$	0.046***	0.028**	0.023	*$SENT.OSC_{t-1}$*	0.055***	0.029**	0.017	*$SENT.OSC_{t-1}$*	0.057***	0.041***	0.001
	[0.011]	[0.012]	[0.019]		[0.010]	[0.012]	[0.016]		[0.013]	[0.012]	[0.015]
CONTROLS	Y	Y	Y	*CONTROLS*	Y	Y	Y	*CONTROLS*	Y	Y	Y
Fixed Effects	Y	Y	Y	Fixed Effects	Y	Y	Y	Fixed Effects	Y	Y	Y
R-squared	0.092	0.093	0.147	R-squared	0.079	0.100	0.129	R-squared	0.111	0.108	0.125
Firms	1799	1766	1628	Firms	1814	1660	1685	Firms	1814	1806	1805
Observations	40,091	38,021	23,488	Observations	47,813	32,963	23,267	Observations	35,872	35,510	32,746

Panel B: Moderating effect of Market Sentiment on Social Media and Crash Risk

	BW Sentiment				*Orthogonalised BW Sentiment*				*HJTZ PLS Sentiment*		
$NCSKEW_t$	*Negative*	*Neutral*	*Positive*	***$NCSKEW_t$***	*Negative*	*Neutral*	*Positive*	***$NCSKEW_t$***	*Negative*	*Neutral*	*Positive*
$SENT.OSC_{t-1}$	0.035***	0.039***	0.056***	*$SENT.OSC_{t-1}$*	0.016	0.032***	0.070***	*$SENT.OSC_{t-1}$*	0.032**	0.053***	0.055***
	[0.011]	[0.012]	[0.013]		[0.012]	[0.011]	[0.013]		[0.014]	[0.013]	[0.014]
CONTROLS	Y	Y	Y	*CONTROLS*	Y	Y	Y	*CONTROLS*	Y	Y	Y
Fixed Effects	Y	Y	Y	Fixed Effects	Y	Y	Y	Fixed Effects	Y	Y	Y
R-squared	0.117	0.109	0.121	R-squared	0.124	0.116	0.113	R-squared	0.138	0.136	0.148
Firms	1811	1813	1812	Firms	1811	1812	1811	Firms	1797	1803	1803
Observations	33,291	35,510	35,257	Observations	34,849	34,664	34,560	Observations	27,322	28,788	29,148

Table 4.9: Impact of lagged Sentiment Oscillations on Crash Risk

Notes: The table presents the regression results of the lagged sentiment oscillations and crash risk. We use ordinary least square (OLS) regression using industry and time fixed effects in Panel A and B. In Panel A, the dependent variable is $\boldsymbol{NCSKEW_t}$, defined as the negative coefficient of the skewness. In Panel B, the dependent variable is $\boldsymbol{DUVOL_t}$, defined as the degree of up and down volatility of the firm-level returns. In both the panels, the independent variable is the sentiment oscillations ($\boldsymbol{SENT.OSC_{t-n}}$) up to n lags, and in this case, the results are presented up to six lags. All continuous variables are winsorized at 1st and 99th percentiles and standardized. *, **, *** represent significance at 10%, 5% and 1% level, respectively. The standard errors are reported in brackets and are robust to heteroscedasticity and clustered at the firm-level.

Panel A

$NCSKEW_t$	*1*	*2*	*3*	*4*	*5*	*6*
$SENT.OSC_{t-1}$	0.040*** [0.006]					
$SENT.OSC_{t-2}$		0.043*** [0.006]				
$SENT.OSC_{t-3}$			0.038*** [0.006]			
$SENT.OSC_{t-4}$				0.038*** [0.006]		
$SENT.OSC_{t-5}$					0.033*** [0.006]	
$SENT.OSC_{t-6}$						0.033*** [0.006]
CONTROLS	Y	Y	Y	Y	Y	Y
Fixed Effects	Y	Y	Y	Y	Y	Y
R-squared	0.052	0.052	0.053	0.054	0.054	0.055
Firms	1815	1813	1813	1812	1811	1810
Observations	104,137	100,326	98,898	97,418	95,871	94,497

Panel B

$DUVOL_t$	*1*	*2*	*3*	*4*	*5*	*6*
$SENT.OSC_{t-1}$	0.036*** [0.007]					
$SENT.OSC_{t-2}$		0.035*** [0.008]				
$SENT.OSC_{t-3}$			0.033*** [0.007]			
$SENT.OSC_{t-4}$				0.029*** [0.007]		
$SENT.OSC_{t-5}$					0.026*** [0.007]	
$SENT.OSC_{t-6}$						0.023*** [0.007]
CONTROLS	Y	Y	Y	Y	Y	Y
Fixed Effects	Y	Y	Y	Y	Y	Y
R-squared	0.064	0.065	0.066	0.067	0.067	0.069
Firms	1814	1813	1812	1810	1810	1810
Observations	72,315	69,764	68,846	67,619	66,616	65,654

Table 4.10: Alternative measures of Sentiment Oscillation and Crash Risk

Notes: The table presents the regression results after using alternative measures of sentiment oscillations and crash risk. In the models where dependent variables are continuous (***$NCSKEW_t$, $DUVOL_t$,*** and ***$NCMRET_t$***), we use ordinary least square regression using industry and time fixed effects. In the model where the dependent variable is dichotomous (***$CRASH309_t$, $CRASH320_t$***, and ***$CRASH20_t$***), we use logistics regression using industry and time fixed effects. Panel A presents the regression results where the independent variable, sentiment oscillations (***$SENT.OSC_{t-1}$***), are predicted using alternative machine learning approaches, i.e., Random Forest Decision Trees (RFDT) and Support Vector Machines (SVM), respectively. For each alternative prediction model, the dependent variables are the ***$NCSKEW_t$*** defined as the negative coefficient of the skewness, ***$DUVOL_t$*** defined as the degree of up and down volatility of the firm-level returns, and ***$CRASH309_t$*** defined as the dummy variable that equals one if the weekly return decline more than 3.09 standard deviation within the given month. Panel B presents the regression results using alternative measures of crash risk. The dependent variables are the ***$NCMRET_t$***, defined as the negative coefficient of the minimum return, ***$CRASH320_t$*** is the dummy variable equals one if the weekly return decline more than 3.20 standard deviation within the given month and, ***$CRASH20_t$*** is the dummy variable equals one if the firm experience at least one market-adjusted weekly return less than 20% within the given month. The independent variable is the sentiment oscillations (***$SENT.OSC_{t-1}$***). All continuous variables are winsorized at 1st and 99th percentiles and standardized. *, **, *** represent significance at 10%, 5% and 1% level, respectively. The standard errors are reported in brackets and are robust to heteroscedasticity and clustered at the firm-level.

Panel A: Alternative prediction measures of Sentiment Oscillations

	Random Forest Decision Trees			***Support Vector Machine***		
	$NCSKEW_t$	***$DUVOL_t$***	***$CRASH309_t$***	***$NCSKEW_t$***	***$DUVOL_t$***	***$CRASH309_t$***
$SENT.OSC_{t-1}$	0.020*** [0.006]	0.018*** [0.007]	0.049* [0.027]	0.033*** [0.006]	0.030*** [0.007]	0.032 [0.028]
CONTROLS	Y	Y	Y	Y	Y	Y
Fixed Effect	Y	Y	Y	Y	Y	Y
R-squared	0.051	0.063		0.051	0.064	
Pseudo R-squared			0.015			0.015
Firms	1815	1814	1815	1815	1814	1815
Observations	104,137	72,315	69,279	104,137	72,315	69,279

Panel B: Alternative measures of Crash Risk

	$NCMRET_t$	***$CRASH320_t$***	***$CRASH20_t$***
$SENT.OSC_{t-1}$	0.037*** [0.011]	0.069*** [0.019]	0.242*** [0.024]
CONTROLS	Y	Y	Y
Fixed Effect	Y	Y	Y
R-squared	0.162		
Pseudo R-squared		0.041	0.139
Firms	1815	1815	1815
Observations	104,137	103,338	102,317

Appendix 4.1: Variables Definitions

Variable	Source	Definition
		Crash Risk
$NCSKEW_t$	CRSP	The third moment of firm-specific weekly returns divided by the weekly returns' standard deviation to the third power multiplied by negative one. The firm-specific weekly returns are estimated based on Equation (9) using 52 weeks rolling regression. Weekly returns are calculated based on five trading days starting from Monday.
$DUVOL_t$		The down-to-up volatility is the log of the ratio of the standard deviations of down weeks and up weeks firm-specific returns multiplied by negative one. The firm-specific weekly returns are estimated based on Equation (9) using 52 weeks rolling regression. Weekly returns are calculated based on five trading days starting from Monday.
$NCMRET_t$		The minimum market-adjusted weekly returns over the most recent 26 weeks are divided by the market-adjusted weekly returns of *t-1* multiplied by negative one. Where the market-adjusted weekly returns are calculated as the difference between the stock returns and CRSP value-weighted returns. Weekly returns are calculated based on five trading days starting from Monday.
$CRASH309_t$		A dummy variable equals one for the given month if the firm's stock encounter at least one firm-specific weekly returns that decline more than 3.09 standard deviations below the average firm-specific weekly returns during the sample period. The firm-specific weekly returns are estimated based on Equation (9) using 52 weeks rolling regression. Weekly returns are calculated based on five trading days starting from Monday.
$CRASH320_t$		A dummy variable equals one for the given month if the firm's stock experiences at least one firm-specific weekly return that falls more than 3.20 standard deviations below the average firm-specific weekly returns during the sample period. The firm-specific weekly returns are estimated based on Equation (9) using 52 weeks rolling regression. Weekly returns are calculated based on five trading days starting from Monday.
$CRASH20_t$		A dummy variable equals one if the firm's market-adjusted weekly returns fall more than 20% in any week within the given month. Where the market-adjusted weekly returns are calculated as the difference between the stock returns and CRSP value-weighted returns. Weekly returns are calculated based on five trading days starting from Monday.
		Sentiment Oscillations
$SENT.OSC_{t-1}$	StockTwits	The log of the 3-months moving average of the sum of the number of times a distinct investor updates her sentiment (*Bullish-to-Bearish* or vice versa) aggregated at the firm-month level. The investors' sentiment is predicted using the Maximum Entropy Method by using self-disclosed sentiment as training data.

Variable	Source	Definition
$SENT.OSC\_Pro_{t-1}$		Sentiment oscillations of *Professional* investors based on nonmissing self-disclosed investors' experience on StockTwits. Whereas the sentiment oscillations are the log of the 3-months moving average of the sum of the number of times a distinct investor updates her sentiment (*Bullish-to-Bearish* or vice versa) aggregated at the firm-month level.
$SENT.OSC\_Novice_{t-1}$		Sentiment oscillations of intermediate and novice (classified as *Novice*) investors based on nonmissing self-disclosed investors' experience on StockTwits. Whereas the sentiment oscillations are the log of the 3-months moving average of the sum of the number of times a distinct investor updates her sentiment (*Bullish-to-Bearish* or vice versa) aggregated at the firm-month level.
$SENT\_OSC\_Ind_{t-1}$		The log of the 3-months moving average of the sum of the number of times a distinct investor updates her sentiment (Bullish-to-Bearish or vice versa) aggregated at the industry-month level. The investors' sentiment is predicted using the Maximum Entropy Method by using self-disclosed sentiment as training data.
		Instrument Variables
$PSYCH\_DIST_{t-1}$	Douglas Dow Website	The psychic distances among investors who are posting ideas about the specific firms in our sample. Since we are using the US sample firms, we use the United States as the base country and then use Dow and Karunaratna's (2006) psychic distance measures using investors' self-disclosed locations on StockTwits at the country-level excluding the United States.
$PROXIMITY_{t-1}$	StockTwits	The distance between self-disclosed investors' location on StockTwits at the city-level within the United States and the city of the firms' headquarter.
		Baseline Controls
$NCSKEW_{t-1}$	CRSP	Lagged value of negative conditional skewness.
$SIGMA_{t-1}$	CRSP	The standard deviation of firm-specific weekly returns within the given month.
RET_{t-1}	CRSP	Average firm-specific weekly returns within the given month.
$DTURN_{t-1}$	CRSP	The detrended turnover is calculated as the average monthly turnover in the last six months, normalized by the average turnover in the last 18 months.
$KURT_{t-1}$	CRSP	The Kurtosis of firm-specific weekly returns within the given month.
$MEDIA_{t-1}$	TRNA	The log of the number of news articles discussing the sample firms in the lead paragraph within the given month. The data is collected from *Thomson Reuters News Analytics (**TRNA**)*.
ROA_{t-1}	Compustat	Quarterly net income at the firm-level is normalized by the total assets.
BTM_{t-1}	Compustat	Quarterly book value of common equity of the firm normalized by the market capitalization.
LEV_{t-1}	Compustat	Quarterly total liabilities of the firm normalized by firms' total assets.

Variable	Source	Definition
$OPACITY_{t-1}$	Compustat	Financial reporting opacity calculated as in Hutton et al. (2009) as the 36-months moving sum of discretionary accruals' absolute value. The discretionary accruals are calculated using the modified Jones (1991) model.
$GOOD_{t-1}$	Compustat	Quarterly value of the Goodwill of the firm normalized by total assets.
$SIZE_{t-1}$	Compustat	The log of quarterly firms' sales.
RD_{t-1}	Compustat	Quarterly firm-level R&D expenditure normalized by the firms' total assets.
AGE10	Compustat	Dummy variable equals one if the firm's age is less than 10 years.
		Additional Variables
$\Delta OPTIONS\_CEO_{t-1}$	Execucomp	The dollar increase in CEO option holdings given a 1% change in underlying stock price following the Core and Guay (2002) yearly approximation reported in thousands.
$\Delta OPTIONS\_CFO_{t-1}$	Execucomp	The dollar increase in CFO option holdings given a 1% change in underlying stock price following the Core and Guay (2002) yearly approximation reported in thousands.
$\Delta EQUITY\_CEO_{t-1}$	Execucomp	The dollar increase in the CEO equity portfolio given a 1% change in underlying stock price following the Core and Guay (2002) yearly approximation reported in thousands.
$\Delta EQUITY\_CFO_{t-1}$	Execucomp	The dollar increase in CFO equity portfolio given 1% change in underlying stock price following the Core and Guay (2002) yearly approximation reported in thousands.
$CSCORE_t$	Compustat	*C Score* is the proxy of conservatism. Following Khan and Watts (2009), *C Scores* are calculated by estimating a firm-level cross-sectional regression between earnings and the firm size, market-to-book ratio, and leverage. *C Score* is a yearly measure.
$ANALYSTS_t$	IBES	The log of the number of analysts covering the sample firms.
$COMPETITION_t$	Compustat	Herfindahl-Hirschman Index (HHI) of firms' quarterly total assets.
$BW.SENT_t$ $BW.SENT\_Orth_t$	BW Website	Following Baker and Wurgler (2006), we use BW sentiment. The data is downloaded from Jeffrey Wurgler's website. Yearly data is converted to monthly data based on the code available on their website.
$PLS.SENT_t$	HJTZ Data Website	Following Huang et al. (2015), we use the partial least sentiment index available on their website. The sentiment data is calculated monthly and available until 2016.

Appendix 4.2: Randomly selected StockTwits' Ideas

Text	Text
I could easily see $BAC go higher, but I have been carrying this burden for six weeks now, and it could easily crash tomorrow. Out at 6.27	$ELLI Haven't seen any news to support this crash. Down $2. unbelievable. makes no sense at all.
$TSLA "tend to personal matters" yeah yeah 3 months before Enron crashed and burned Jeffrey Skilling resigned to pursue personal interests..	$FSLR days/moves like this, remind me to never forget the flash crash and the capabilities of computers trading in the markets
$IDIX when it eventually crashes, no one will even say a word. When it turns up again ill be long because it moves hard, but in the end...	$BBY another upgrade. more vertical it goes, the sharper the crash. last Quarter didn't give guidance. Q4's surprise was one hit wonder.
Get ready for the posts. $AAPL is off the high. Looks tired to me. In fact, I think it is going to crash. Probably finish the day red. LOL	$MBI Based on the PE and EPS this looks like a cheap stock. Am I missing something here? Was at 70 prior to the crash. What's the deal?
$DMND Going lower today. Yesterdays close all but sealed that. Waiting to see what happens at 20.50 which is essentially the post crash low.	$SHLD might I say about damn time for some rain in the desert! Been holding this since the Nov crash.
$FIO Short convince themselves expiration will crash stock, then cover like mad when it fails to materialize	$SCTY Isn't this supposed to show an ER loss? Confused why this isn't crashing. Is it going to spike tomorrow even after the loss??
Good call...$AAPL is not a silver my friend. $AAPL is better than GOLD. Everyone waiting to crash but it won't happen..Buy $AAPL	$ARIA The worst is when it was 4.5 and crash twice back down into the 2's. Those are painful memories. I like it better this way.
$TNGO down on big volume and crashed through 200 day EMA. No reason to try any long trade for now. stockcharts.com/h-sc/ui?s=T...	$TWTR These momo stocks need to crash to get a proper clearout of this market. They are so overvalued they could wipe plenty of the indexes
$KCG Will be green all day today and close green. The 30 days since the crash will be up before the next session. Just bought more.	$CMG I think this is way over valued. The earning is close, but any weakness will crash it hard. I have been short since the last earning!!
$SKUL call me optimistic, but this crash should signal shorts to start covering	$VNDA The only way they can withstand this price is with any news, if not this will crash. Honestly if u made a profit,take it
Institutions that issued $VRTX may calls will want to crash this as close to 60 as possible this week to minimize millions of call losses...	$JCP Funds will crash this tomorrow, as they will focus on the bad numbers, and guidance will confuse.
$TSLA Its pretty far up here, gunna have to really hit a hr w/ the earnings to not crash...im on the sidelines	$WRLD so reading through the financials, they really increased lending over the last year! Setting up for a complete crash imo...
$GTAT any neg mention tonight here on aapl call and this stock will crash to 12 ish..where id buy 1/2 back out of all at 17.20	$MDVN Given the massive crash I'm act. pleased with MDVN (though I wish it'd gone lower for my buy). Down 10-20% is overperform right now.

Appendix 4.3: Randomly selected StockTwits ideas discussing CEO/CFO options trading

Text	Text
Cashing out +400%, his new 2012 stock options coming RT @Tradeday $VVUS CFO bailed on 329k shares..tell u something?	$BBY Best Buy CEO Joly sells 100,686 shares and 350,467 stock options - nice sell into momentum.
Dr @JasonMilano8 The ceo of $KERX is buying all of the top staff is buying The CFO might have optioned at 1.60 but We are all in with you.	Nike's EVP & CFO just cashed-in 33,000 options $NKE
$LQDT It's the CEO who's been selling off shares; the CFO merely exercised options and bought them back and it wasn't a large amount.	Financial Engines's EVP and CFO just cashed-in 5,000 options $FNGN
President of $COO (John Weber) doing same thing as CEO. Exercising stock options, which were most likely pre-planned.	Wright Medical Group's Sr. VP & CFO just cashed-in 144 options $WMGI
$QCOR CEO profited $11.9 million from exercising QCOR options in 2012. More active at trading than other CEO's I've seen.	$RMTI VP-CFO stock option(right to buy) exercised
$BAC CEO Did'nt get a raise. He was awarded more stock options for job performance. Media carries false story to demonize CEO's and banks.	$C Form 4: On May 21, 2015, Francisco Aristeguieta, CEO, Latin America, exercised 16,951.7 stock options and acquired the same at $ $40.8.
CEO and CFO of $SGI exercised options (in predetermined agreement). Bought back what they sold.	$CSC Form 4: Paul Saleh, CFO exercised 35,000 stock options on June 15, 2015 owning 118,146 shares at that time.
@CalD @Burp $UNXL don't forget about all the options / warrants that the CFO also owns. Over 2 million for whole company according to 10-q	The CFO of Mettler-Toledo is Exercising Options $MTD

Chapter 5

5. Conclusion

5.1. Summary of Research Findings and Contribution

Social media has recently emerged as a powerful tool to facilitate investors to share their analyses and recommendations. In this thesis, using novel datasets from Twitter and StockTwits, I examine the association between investor-oriented social media platforms and the financial markets. This thesis contributes to the emerging literature of social finance and sheds light on the significance of social media platforms in the financial markets.

The thesis is organized in the form of three related but self-contained chapters. The second chapter focuses on investors' attention allocation while discussing firms on Twitter using cashtags. In this chapter, I examine the impact of attention allocation to predict investors' trading behaviors in the financial markets. I find that social media attention (SMA) on Twitter can only predict retail investors' trading behavior, i.e., an increase in SMA predicts an increase in retail investors' net order flow. This result suggests the transactional role of social media, facilitating investors by providing the opportunity to share their opinions and consume information and, consequently, increasing the velocity of information diffusion in the financial markets. Moreover, the impact of SMA is more pronounced when tweets are posted by verified users, a single tweet generates long discussion threads, and a tweet is retweeted several times. In contrast, we find no significant association between SMA and institutional investors' trades.

This chapter contributes to the emerging literature on social media's role in the financial markets. It provides unique evidence on how investors allocate their attention in the financial markets, which further leads us to another vital question: Which segment of the market is represented by these social media platforms? The data analysis provides evidence that large social networks like Twitter mainly attract retail investors. Furthermore, the findings suggest that retail investors play a pivotal role in increasing the velocity of information in the financial markets. The main contribution of this chapter is that it highlights the growing role of social media in financial markets and invites researchers to investigate further how such platforms influence investors' opinions.

Building upon these arguments, the third chapter investigates what happens once investors allocate their attention by posting ideas and sharing their opinions on social media platforms. In this chapter, I investigate the consequences of attention allocation on social media platforms for investors. Using a novel dataset from an investor-oriented social media platform, i.e., StockTwits, I examine whether disagreement among investors on StockTwits increases the flow of firm-specific information in the financial markets? The analysis reveals a negative association between disagreement on StockTwits and stock return synchronicity, suggesting that increase in disagreement among investors on StockTwits increases the flow of firm-specific information in the financial markets. Further tests reveal that disagreement among investors increases the stock price informativeness by increasing the price leads of earnings.

To reaffirm the research findings, I have used two additional tests. First, by examining the moderating effect of media coverage, I find that StockTwits acts as a catalyst when there is no/low media coverage by providing firm-specific information. It acts as an intermediary and exacerbates firm-specific information flows into financial markets when there is greater media coverage. Second, by examining the interaction between recommendation revisions and disagreement, I find that updates in investors' recommendations pronounce the impact of disagreement, suggesting that investors update their analysis based on the increase in the inflow of firm-specific information. These results remain valid after controlling for endogeneity using the instrumental variable approach and self-selection bias using the Heckman selection model.

The results of this chapter generate some valuable insights, as I examine the association between disagreement further and return synchronicity. First, the firm information environment analysis indicates that disagreement increases for opaque firms, firms with greater diversity and industry concentration, and firms with more insider trades. Second, One of the key characteristics of social media is its attention-grabbing features, i.e., the salience of information signals. Therefore, the analysis of salience of information signals presents that the increase in the salience of information signals on StockTwits increases disagreement, i.e., the higher inflow of firm-specific information.

This chapter contributes to the existing literature on behavioral finance in several ways. First, it provides an essential piece of the puzzle by explaining that the social media platform for investors can predict financial markets because it allows firm-specific information to flow to investors who actively participate in discussions on such platforms and update their priors based on available information. Second, it highlights the significance of social media platforms for investors by providing evidence that disagreement increases for firms with less transparent information environments. Thus, facilitating investors to consume firm-specific information from alternative channels. Third, it contributes to the emerging literature on the role of the salience of information signals in financial markets by highlighting the significance of information diffusion and heterogeneity of investors in a large social network. Finally, it contributes to the academic debate on whether stock prices for firms with a low R^2 statistics are more informative by providing substantial evidence that less synchronicity can reflect higher stock price informativeness.

In chapter four, I further consolidate my research findings by examining the investors' efficiency on StockTwits to predict firm-level future crash risk. I accomplish this by developing sentiment oscillations, a new proxy of change in investors' sentiment weighted by the number of followers each investor has on StockTwits. I find that sentiment oscillations on StockTwits can predict firm-level future crash risk. The results in this chapter suggest that social interaction among investors on StockTwits facilitates these investors to consume firm-specific information, which enables them to anticipate such firm-specific future events. More importantly, the large social network of investors increases the velocity of information diffusion of such firm-specific information, causing a contagion effect in the financial markets.

I employ two mechanisms to validate such research findings. First, prior literature on crash risk suggests that crashes occur when managers conceal bad news and earnings announcement events reflect such information (Kothari et al., 2009). Furthermore, Ak et al. (2016) find that more than 70% of crashes occur due to earnings announcements. This chapter finds a strong association between sentiment oscillations with future crash risk during earnings announcement months and a weak association during non-earnings announcement months. This result shows that investors on StockTwits can anticipate bad news horading by the managers, which motivates them to interact with their peers and update their sentiment.

Second, Hong and Stein (1999) and Hong and Stein (2007) highlight the role of market segmentation and specializations by suggesting that some investors in the financial markets act as a frontrunner to access value-relevant information. Motivated by these studies, I examine how investors' heterogeneity accentuates the power of sentiment oscillations to predict firm-level future crash risk. These results suggest that professional investors have access to more information sources than novice investors. Therefore, sentiment oscillations of professional investors lead the market and predict firm-level future crash risk.

In this chapter, I employ various tests to explore the monitory role of investors on StockTwits. First, using CEO and CFO stock- and options-based compensation data, I find that investors on StockTwits closely monitor CEO and CFO options portfolios. However, only the CFO options portfolio pronounce the impact of sentiment oscillations to predict firm-level future crash risk. These findings are consistent with Kim et al. (2011b), who present evidence that CFO option incentives increase the crash risk. In the case of investors on StockTwits, they update their sentiment depending on the CFO options trading. Second, measuring firm-level financial reporting conservatism, analysts coverage, and competition, I find that the impact of sentiment oscillations is more pronounced for firms with less conservatism, less analyst coverage, and decreased product market competition. This suggests that investors on StockTwits analyze the available information; however, for firms with less firm-specific information, they overcome this deficiency by providing firm-specific information based on their analysis and updating their sentiment.

Chapter four contributes to the extant literature in a number of ways. First, despite the significance of crash risk, there is a lack of evidence to investigate the role of investors in the financial markets to predict the firm-level future crash risk. This study provides firsthand evidence suggesting that investors update their sentiment based on their social interaction with other investors and the available information signals. These updates empower investors to predict the firm-level future crash risk. Second, it presents empirical evidence on the role of investors' heterogeneity and information diffusion via a large social network of investors in the financial markets and how these investors' specific attributes enable them to predict firm-level future crash risk. Third, it highlights the significance of social media platforms by providing evidence that such platforms can be used as an analytical tool by investors to predict firm-level future events.

5.2. Practical Implications

The research findings in this thesis have practical implications for firms, regulators, investors, and professional money managers and present substantial evidence that social media platforms for investors have emerged as a powerful tool that can diffuse information faster than any other medium of information.

Firms can use social media platforms to release any firm-specific information. More importantly, such platforms can decrease the information distribution cost for firms to reach a wider audience than traditional information distribution channels. Furthermore, managers can use such platforms to entice investors to invest in their stock and build their reputation. Recently, there is an increasing trend that managers have started using such platforms to promote their firms and stay connected with their investors and other stakeholders to improve their firms' reputations. In contrast, social media platforms have also emerged as a powerful tool that can influence investors' opinions and affect firms' and managers' reputations.

Social media platforms have often drawn the regulators' attention and established their significance. Overall, information sharing on social media platforms is not regulated by the Securities and Exchange Commission, USA. This study provides corroborative evidence that social interactions among investors increase the inflow of firm-specific information. Investors on such platforms can predict firm-level future crash risk. Throughout this study, I find consistent evidence that investors follow other influential investors, learn from their analysis, and update their sentiment. Overall, this study draws regulators' attention to regulate such platforms. This will benefit all the stakeholders, reduce information asymmetry and increase the efficiency of the stock prices.

For investors, the research findings may encourage them to share their ideas, analysis, and recommendations. This not only reduces the information asymmetry between investors and the firms but also facilitates investors to access firm-specific information at low/no cost. More importantly, the fourth chapter highlights the benefit of using social media as a tool to predict stock price crashes, which can equally benefit investors and other stakeholders.

For professional money managers, the research findings can motivate them to use social media as an analytical tool to facilitate investors and diversify their portfolios' risk. More specifically, the research findings of the third chapter verify that social media platforms provide firm-specific information. In chapter four, I provide firsthand evidence that market participants can anticipate stock price crashes by using information from such platforms and can consequently adjust their market positions and derive benefits. In this regard, Bloomberg and Thomson Reuters have already started offering social media feeds to their subscribers.

5.3. Research Limitations

It is often the research limitations that generate opportunities for future research. In this study as well, some of the limitations being discussed in this section can encourage researchers to conduct further studies to examine the role of social media. In the second chapter, I use one of the most extensive firm-level social media datasets from Twitter. However, due to the inherent nature of the Twitter application programming interface (API), it limits access to tweets and other related information. Therefore, during the data collection process, I could not collect all the tweets discussing sample firms, resulting in sample selection bias. Moreover, I do not deploy any fake news detection mechanism as it is beyond the scope of this study. However, fake news can result in increased activity on Twitter. Finally, although I employ several tests to clean the Twitter data, the subjective nature of the tweets poses a significant challenge. Therefore, some of the tweets in my dataset in the second chapter can add noise to the existing sample of tweets.

I overcome these limitations in subsequent chapters by using an investor-oriented social media platform, i.e., StockTwits. As part of my research methodology in these chapters, I use three alternative machine learning models to predict investors' sentiment on StockTwits. However, using advanced machine learning models may improve prediction accuracy, which may influence these chapters' results.

Despite these limitations and constraints in data collection, the research findings in this thesis contribute extensively to the emerging field of social finance and highlight the significance of social media platforms for investors.

5.4. Suggestions for Future Research

The research in this thesis encourages scholars to examine the role of social media platforms in the financial markets in the following areas. First, although in the second chapter, I assume that social bots cannot influence investors' opinions and their information is subject to the crowding-out effect in the financial markets. It is imperative to examine the influence of social bots in the financial markets. Second, while the research findings in chapter 3 highlight the critical role of disagreement among investors, it will be insightful to examine investors' heterogeneity on StockTwits based on their investment techniques. Finally, the research findings in chapter 4 suggest that besides crash risk predictions, it would be insightful to investigate social media's role in creating market frenzies and other firm-specific events.

References

Acemoglu, D., Chernozhukov, V., & Yildiz, M. (2006). Learning and disagreement in an uncertain world. 1-51. doi:10.2139/ssrn.939169

Ahern, K. R., & Sosyura, D. (2014). Who writes the news? Corporate press releases during merger negotiations. *The Journal of Finance, 69*(1), 241-291. doi:10.1111/jofi.12109

Ahern, K. R., & Sosyura, D. (2015). Rumor has it: Sensationalism in financial media. *The Review of Financial Studies, 28*(7), 2050-2093. doi:10.1093/rfs/hhv006

Ak, B. K., Rossi, S., Sloan, R., & Tracy, S. (2016). Navigating stock price crashes. *The Journal of Portfolio Management, 42*(4), 28-37. doi:10.3905/jpm.2016.42.4.028

Al-Nasseri, A., & Menla, F. A. (2018). What does investors' online divergence of opinion tell us about stock returns and trading volume? *Journal of Business Research, 86*, 166-178. doi:10.1016/j.jbusres.2018.01.006

Al Mamun, M., Balachandran, B., & Duong, H. N. (2020). Powerful CEOs and stock price crash risk. *Journal of Corporate Finance, 62*, 101582. doi:10.1016/j.jcorpfin.2020.101582

Ali, A., Klasa, S., & Yeung, E. (2009). The limitations of industry concentration measures constructed with compustat data: Implications for finance research. *The Review of Financial Studies, 22*(10), 3839-3871. doi:10.1093/rfs/hhn103

Ali, A., Klasa, S., & Yeung, E. (2014). Industry concentration and corporate disclosure policy. *Journal of accounting and economics, 58*(2-3), 240-264. doi:10.1016/j.jacceco.2014.08.004

An, H., & Zhang, T. (2013). Stock price synchronicity, crash risk, and institutional investors. *Journal of Corporate Finance, 21*, 1-15. doi:10.1016/j.jcorpfin.2013.01.001

An, Z., Chen, C., Naiker, V., & Wang, J. (2020). Does media coverage deter firms from withholding bad news? Evidence from stock price crash risk. *Journal of Corporate Finance, 64*, 101664. doi:10.1016/j.jcorpfin.2020.101664

Andreou, C. K., Andreou, P. C., & Lambertides, N. (2021). Financial distress risk and stock price crashes. *Journal of Corporate Finance, 67*, 101870. doi:10.1016/j.jcorpfin.2020.101870

Andreou, P. C., Louca, C., & Petrou, A. P. (2017). CEO age and stock price crash risk. *Review of Finance, 21*(3), 1287-1325. doi:10.1093/rof/rfw056

Antoniou, C., Doukas, J. A., & Subrahmanyam, A. (2013). Cognitive dissonance, sentiment, and momentum. *Journal of Financial and Quantitative Analysis, 48*(1), 245-275. doi:10.1017/S0022109012000592

Antoniou, C., Doukas, J. A., & Subrahmanyam, A. (2016). Investor sentiment, beta, and the cost of equity capital. *Management Science, 62*(2), 347-367. doi:10.1287/mnsc.2014.2101

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance, 59*(3), 1259-1294. doi:10.1111/j.1540-6261.2004.00662.x

Antweiler, W., & Frank, M. Z. (2006). Do US stock markets typically overreact to corporate news stories? *SSRN Research Paper No. 878091*. doi:10.2139/ssrn.878091

Antón, M., & Polk, C. (2014). Connected stocks. *The Journal of Finance, 69*(3), 1099-1127. doi:10.1111/jofi.12149

Armstrong, C. S., Guay, W. R., & Weber, J. P. (2010). The role of information and financial reporting in corporate governance and debt contracting. *Journal of accounting and economics, 50*(2-3), 179-234. doi:10.1016/j.jacceco.2010.10.001

Aumann, R. J. (1976). Agreeing to disagree. *The Annals of Statistics 4*(6), 1236-1239. doi:10.1214/aos/1176343654

Ayers, B. C., & Freeman, R. N. (2003). Evidence that analyst following and institutional ownership accelerate the pricing of future earnings. *Review of Accounting Studies, 8*(1), 47-67. doi:10.1023/A:1022647822683

Ayotte, K. (2020). Disagreement and capital structure complexity. *The Journal of Legal Studies, 49*(1), 1-32. doi:10.1086/708671

Azar, P. D., & Lo, A. W. (2016). The wisdom of Twitter crowds: Predicting stock market reactions to FOMC meetings via Twitter feeds. *The Journal of Portfolio Management, 42*(5), 123-134. doi:10.3905/jpm.2016.42.5.123

Baik, B., Cao, Q., Choi, S., & Kim, J.-M. (2016). Local Twitter activity and stock returns. *SSRN Research Paper No. 2783670*. doi:10.2139/ssrn.2783670

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance, 61*(4), 1645-1680. doi:10.1111/j.1540-6261.2006.00885.x

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of economic perspectives, 21*(2), 129-151. doi:10.1257/jep.21.2.129

Baker, M., & Wurgler, J. (2013). Behavioral corporate finance: An updated survey. In E. Eckbo (Ed.), *Handbook of the Economics of Finance* (Vol. 2, pp. 357-424): Elsevier.

Ball, R., Jayaraman, S., & Shivakumar, L. (2012). Audited financial reporting and voluntary disclosure as complements: A test of the confirmation hypothesis. *Journal of accounting and economics, 53*(1-2), 136-166. doi:10.1016/j.jacceco.2011.11.005

Baloria, V. P., & Heese, J. (2018). The effects of media slant on firm behavior. *Journal of Financial Economics, 129*(1), 184-202. doi:10.1016/j.jfineco.2018.04.004

Bamber, L. S., Barron, O. E., & Stober, T. L. (1999). Differential interpretations and trading volume. *Journal of Financial and Quantitative Analysis, 34*(3), 369-386. doi:10.2307/2676264

Banerjee, S., Davis, J., & Gondhi, N. (2019). Choosing to disagree in financial markets. *SSRN Research Paper No. 3335257*. doi:10.2139/ssrn.3335257

Banerjee, S., Kaniel, R., & Kremer, I. (2009). Price drift as an outcome of differences in higher-order beliefs. *The Review of Financial Studies, 22*(9), 3707-3734. doi:10.1093/rfs/hhp014

Banerjee, S., & Kremer, I. (2010). Disagreement and learning: Dynamic patterns of trade. *The Journal of Finance, 65*(4), 1269-1302. doi:10.1111/j.1540-6261.2010.01570.x

Barber, & Odean, T. (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The Review of Financial Studies, 21*(2), 785-818. doi:10.1093/rfs/hhm079

Barber, Odean, T., & Zhu, N. (2008). Do retail trades move markets? *The Review of Financial Studies, 22*(1), 151-186. doi:10.1093/rfs/hhn035

Barber, B. M., & Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *The Journal of Finance, 55*(2), 773-806. doi:10.1111/0022-1082.00226

Barberis, N. (2018). Psychology-based models of asset prices and trading volume. In *Handbook of Behavioral Economics: Applications and Foundations 1* (Vol. 1, pp. 79-175): Elsevier.

Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics, 49*(3), 307-343. doi:10.1016/s0304-405x(98)00027-0

Barberis, N., Shleifer, A., & Wurgler, J. (2005). Comovement. *Journal of Financial Economics, 75*(2), 283-317. doi:10.1016/j.jfineco.2004.04.003

Barberis, N., & Thaler, R. (2003). *A survey of behavioral finance* (Vol. 1).

Basu, S. (1997). The conservatism principle and the asymmetric timeliness of earnings. *Journal of accounting and economics, 24*(1), 3-37. doi:10.1016/S0165-4101(97)00014-1

Baziotis, C., Pelekis, N., & Doulkeridis, C. (2017, August 3-4, 2017). *Datastories at semeval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis.* Paper presented at the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, Canada.

Bebchuk, L. A. (2009). *Compensation structure and systemic risk*. Washington: U.S Government Printing Office

Becker, K., & Nobre, H. (2013). *Do Firms Manage Their Social Network Reputation? An International Facebook and Twitter Study*. Paper presented at the 5th Annual Euromed Conference of the Euromed Academy of Business: Building New Business Models for

Success through Competitiveness and Responsibility. <Go to ISI>://WOS:000320606900011

Ben-Rephael, A., Da, Z., & Israelsen, R. D. (2017). It depends on where you search: institutional investor attention and underreaction to news. *The Review of Financial Studies, 30*(9), 3009-3047. doi:10.1093/rfs/hhx031

Benmelech, E., Kandel, E., & Veronesi, P. (2010). Stock-based compensation and CEO (dis) incentives. *The Quarterly Journal of Economics, 125*(4), 1769-1820. doi:10.1162/qjec.2010.125.4.1769

Bhagwat, V., & Burch, T. R. (2016). Pump it up? Tweeting to manage investor attention to earnings news. *SSRN Research Paper No. 2382962*. doi:10.2139/ssrn.2382962

Blankespoor, E. (2018). Firm communication and investor response: A framework and discussion integrating social media. *Accounting, Organizations and Society, 68*, 80-87. doi:10.1016/j.aos.2018.03.009

Blankespoor, E., Miller, G. S., & White, H. D. (2014). The role of dissemination in market liquidity: Evidence from firms' use of Twitter. *The Accounting Review, 89*(1), 79-112. doi:10.2308/accr-50576

Bodnaruk, A. (2009). Proximity always matters: Local bias when the set of local companies changes. *Review of Finance, 13*(4), 629-656. doi:10.1093/rof/rfp014

Boehmer, E., Jones, C. M., Zhang, X., & Zhang, X. (2017). Tracking retail investor activity. *The Journal of Finance*.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1-8. doi:10.1016/j.jocs.2010.12.007

Bonsón, E., & Flores, F. (2011). Social media and corporate dialogue: The response of global financial institutions. *Online information review, 35*(1), 34-49. doi:10.1108/14684521111113579

Bordalo, P., Gennaioli, N., Porta, R. L., & Shleifer, A. (2019). Diagnostic expectations and stock returns. *The Journal of Finance, 74*(6), 2839-2874. doi:10.1111/jofi.12833

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association, 90*(430), 443-450. doi:10.1080/01621459.1995.10476536

Bovet, A., & Makse, H. A. (2019). Influence of fake news in twitter during the 2016 US presidential election. *Nature communications, 10*(7), 1-14. doi:10.1038/s41467-018-07761-2

Brockman, P., Liebenberg, I., & Schutte, M. G. (2010). Comovement, information production, and the business cycle. *Journal of Financial Economics, 97*(1), 107-129. doi:10.1016/j.jfineco.2010.03.008

Bushman, R. M., Chen, Q., Engel, E., & Smith, A. (2004). Financial accounting information, organizational complexity and corporate governance systems. *Journal of accounting and economics, 37*(2), 167-201. doi:10.1016/j.jacceco.2003.09.005

Bushman, R. M., Piotroski, J. D., & Smith, A. J. (2004). What determines corporate transparency? *Journal of Accounting Research, 42*(2), 207-252. doi:10.1111/j.1475-679X.2004.00136.x

Callen, J. L., & Fang, X. (2015a). Religion and stock price crash risk. *Journal of Financial and Quantitative Analysis, 50*(1-2), 169-195. doi:10.1017/S0022109015000046

Callen, J. L., & Fang, X. (2015b). Short interest and stock price crash risk. *Journal of Banking & Finance, 60*, 181-194. doi:10.1016/j.jbankfin.2015.08.009

Campbell, J. L., DeAngelis, M. D., & Moon, J. R. (2019). Skin in the game: Personal stock holdings and investors' response to stock analysis on social media. *Review of Accounting Studies, 24*(3), 731-779. doi:10.1007/s11142-019-09498-9

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance, 52*(1), 57-82. doi:10.1111/j.1540-6261.1997.tb03808.x

Carlin, B. I., Longstaff, F. A., & Matoba, K. (2014). Disagreement and asset prices. *Journal of Financial Economics, 114*(2), 226-238. doi:10.1016/j.jfineco.2014.06.007

Cen, L., Wei, K. J., & Yang, L. (2017). Disagreement, underreaction, and stock returns. *Management Science, 63*(4), 1214-1231. doi:10.1287/mnsc.2015.2405

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). *Measuring user influence in twitter: The million follower fallacy.* Paper presented at the Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.

Chahine, S., Mansi, S., & Mazboudi, M. (2015). Media news and earnings management prior to equity offerings. *Journal of Corporate Finance, 35*, 177-195. doi:10.1016/j.jcorpfin.2015.09.002

Chakrabarty, B., Li, B., Nguyen, V., & Van Ness, R. A. (2007). Trade classification algorithms for electronic communications network trades. *Journal of Banking & Finance, 31*(12), 3806-3821. doi:10.1016/j.jbankfin.2007.03.003

Chakrabarty, B., Moulton, P. C., & Shkilko, A. (2012). Short sales, long sales, and the Lee–Ready trade classification algorithm revisited. *Journal of Financial Markets, 15*(4), 467-491. doi:10.1016/J.FINMAR.2012.01.001

Chakrabarty, B., Pascual, R., & Shkilko, A. (2015). Evaluating trade classification algorithms: Bulk volume classification versus the tick rule and the Lee-Ready algorithm. *Journal of Financial Markets, 25*, 52-79. doi:10.2139/ssrn.2182819

Chan, K., & Hameed, A. (2006). Stock price synchronicity and analyst coverage in emerging markets. *Journal of Financial Economics, 80*(1), 115-147. doi:10.1016/j.jfineco.2005.03.010

Chang, S.-L., Hwang, L.-J., & Li, C. A. (2020). The Effect of Managerial Overconfidence on Conference Calls. *The International Journal of Business and Finance Research, 14*(1), 1-12.

Chang, X., Chen, Y., & Zolotoy, L. (2017). Stock liquidity and stock price crash risk. *Journal of Financial and Quantitative Analysis, 52*(4), 1605-1637. doi:10.1017/S0022109017000473

Chava, S., & Purnanandam, A. (2010). CEOs versus CFOs: Incentives and corporate policies. *Journal of Financial Economics, 97*(2), 263-278. doi:10.1016/j.jfineco.2010.03.018

Chelley-Steeley, P., Lambertides, N., & Savva, C. S. (2019). Sentiment, order imbalance, and co-movement: An examination of shocks to retail and institutional trading activity. *European Financial Management, 25*(1), 116-159. doi:10.1111/eufm.12146

Chemmanur, T. J., & Yan, A. (2019). Advertising, attention, and stock returns. *Quarterly Journal of Finance, 9*(3), 1-51. doi:10.1142/S2010139219500095

Chen, H., De, P., Hu, Y., & Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies, 27*(5), 1367-1403. doi:10.1093/rfs/hhu001

Chen, H., Hwang, B.-H., & Liu, B. (2019). The Emergence of'Social Executives' and Its Consequences for Financial Markets. *SSRN Research Paper No. 2318094*. doi:10.2139/ssrn.2318094

Chen, H., Singal, V., & Whitelaw, R. F. (2016). Comovement revisited. *Journal of Financial Economics, 121*(3), 624-644. doi:10.1016/j.jfineco.2016.05.007

Chen, J., Hong, H., & Stein, J. C. (2001). Forecasting crashes: Trading volume, past returns, and conditional skewness in stock prices. *Journal of Financial Economics, 61*(3), 345-381. doi:10.1016/S0304-405x(01)00066-6

Chen, J., Hong, H., & Stein, J. C. (2002). Breadth of ownership and stock returns. *Journal of Financial Economics, 66*(2-3), 171-205. doi:10.1016/S0304-405X(02)00223-4

Chen, Y., Xie, Y., You, H., & Zhang, Y. (2018). Does crackdown on corruption reduce stock price crash risk? Evidence from China. *Journal of Corporate Finance, 51*, 125-141. doi:10.1016/j.jcorpfin.2018.05.005

Chordia, T., Hu, J., Subrahmanyam, A., & Tong, Q. (2017). Order flow volatility and equity costs of capital. *Management Science, 65*(4), 1520-1551. doi:10.1287/mnsc.2017.2848

Chordia, T., & Subrahmanyam, A. (2004). Order imbalance and individual stock returns: Theory and evidence. *Journal of Financial Economics, 72*(3), 485-518. doi:10.1016/S0304-405X(03)00175-2

Chue, T. K., Gul, F. A., & Mian, G. M. (2019). Aggregate investor sentiment and stock return synchronicity. *Journal of Banking & Finance, 108*, 1-19. doi:10.1016/j.jbankfin.2019.105628

Clarke, J., Chen, H., Du, D., & Hu, Y. J. (2020). Fake news, investor attention, and market reaction. *Information Systems Research, 32*(1), 35-52. doi:10.1287/isre.2019.0910

Coles, J. L., Daniel, N. D., & Naveen, L. (2006). Managerial incentives and risk-taking. *Journal of Financial Economics, 79*(2), 431-468. doi:10.1016/j.jfineco.2004.09.004

Cookson, J. A., & Niessner, M. (2019). Why don't we agree? Evidence from a social network of investors. *The Journal of Finance, 75*(1), 173-228. doi:10.1111/jofi.12852

Core, J., & Guay, W. (2002). Estimating the value of employee stock option portfolios and their sensitivities to price and volatility. *Journal of Accounting Research, 40*(3), 613-630. doi:10.1111/1475-679X.00064

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *The Journal of Finance, 66*(5), 1461-1499. doi:10.1111/j.1540-6261.2011.01679.x

Da, Z., Engelberg, J., & Gao, P. (2014). The sum of all FEARS investor sentiment and asset prices. *The Review of Financial Studies, 28*(1), 1-32. doi:10.1093/rfs/hhu072

Da, Z., Huang, X., & Jin, L. J. (2020). Extrapolative beliefs in the cross-section: What can we learn from the crowds? *Journal of Financial Economics, 140*(1), 175-196. doi:10.1016/j.jfineco.2020.10.003

Dang, T. L., Dang, M., Hoang, L., Nguyen, L., & Phan, H. L. (2020). Media coverage and stock price synchronicity. *International Review of Financial Analysis, 67*, 1-18. doi:10.1016/j.irfa.2019.101430

Dang, V. A., Lee, E., Liu, Y., & Zeng, C. (2018). Corporate debt maturity and stock price crash risk. *European Financial Management, 24*(3), 451-484. doi:10.1111/eufm.12134

Daniel, K., Hirshleifer, D., & Subrahmanyam, A. (1998). Investor psychology and security market under- and overreactions. *The Journal of Finance, 53*(6), 1839-1885. doi:10.1111/0022-1082.00077

Darrough, M. N., & Stoughton, N. M. (1990). Financial disclosure policy in an entry game. *Journal of accounting and economics, 12*(1-3), 219-243.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science, 53*(9), 1375-1388. doi:10.1287/mnsc.1070.0704

De Bondt, W. F., & Thaler, R. (1985). Does the stock market overreact? *The Journal of Finance, 40*(3), 793-805. doi:10.1111/j.1540-6261.1985.tb05004.x

Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1995). Detecting earnings management. *The Accounting Review, 70*(2), 193-225.

Diether, K. B., Malloy, C. J., & Scherbina, A. (2002). Differences of opinion and the cross section of stock returns. *The Journal of Finance, 57*(5), 2113-2141. doi:10.1111/0022-1082.00490

Dimson, E. (1979). Risk measurement when shares are subject to infrequent trading. *Journal of Financial Economics, 7*(2), 197-226. doi:10.1016/0304-405X(79)90013-8

Ding, R., Zhou, H., & Li, Y. (2019). Social media, financial reporting opacity, and return comovement: Evidence from Seeking Alpha. *Journal of Financial Markets, 50*, 1-25. doi:10.1016/j.finmar.2019.100511

Dow, D., & Karunaratna, A. (2006). Developing a multidimensional instrument to measure psychic distance stimuli. *Journal of international business studies, 37*(5), 578-602. doi:10.1057/palgrave.jibs.8400221

Drake, M. S., Jennings, J., Roulstone, D. T., & Thornock, J. R. (2017). The comovement of investor attention. *Management Science, 63*(9), 2847-2867. doi:10.1287/mnsc.2016.2477

Durnev, A., Morck, R., & Yeung, B. (2004). Value-enhancing capital budgeting and firm-specific stock return variation. *The Journal of Finance, 59*(1), 65-105. doi:10.1111/j.1540-6261.2004.00627.x

Durnev, A., Morck, R., Yeung, B., & Zarowin, P. (2003). Does greater firm-specific return variation mean more or less informed stock pricing? *Journal of Accounting Research, 41*(5), 797-836. doi:10.1046/j.1475-679X.2003.00124.x

Dyck, A., Morse, A., & Zingales, L. (2010). Who blows the whistle on corporate fraud? *The Journal of Finance, 65*(6), 2213-2253. doi:10.1111/j.1540-6261.2010.01614.x

Dyck, A., Moss, D., & Zingales, L. (2013). Media versus special interests. *Journal of Law & Economics, 56*(3), 521-553. doi:10.1086/673216

Dyck, A., Volchkova, N., & Zingales, L. (2008). The corporate governance role of the media: Evidence from Russia. *The Journal of Finance, 63*(3), 1093-1135. doi:10.1111/j.1540-6261.2008.01353.x

Easley, D., de Prado, M. L., & O'Hara, M. (2016). Discerning information from trade data. *Journal of Financial Economics, 120*(2), 269-285. doi:10.2139/ssrn.1989555

Edmans, A., Garcia, D., & Norli, Ø. (2007). Sports sentiment and stock returns. *The Journal of Finance, 62*(4), 1967-1998. doi:10.1111/j.1540-6261.2007.01262.x

Ellis, K., Michaely, R., & O'Hara, M. (2000). The accuracy of trade classification rules: Evidence from Nasdaq. *Journal of Financial and Quantitative Analysis, 35*(4), 529-551. doi:10.1017/S0022109000014964

Engelberg, J., Sasseville, C., & Williams, J. (2012). Market madness? The case of mad money. *Management Science, 58*(2), 351-364. doi:10.1287/mnsc.1100.1290

Evans, L., Owda, M., Crockett, K., & Vilas, A. F. (2019). A methodology for the resolution of cashtag collisions on Twitter–A natural language processing & data fusion approach. *Expert Systems With Applications, 127*, 353-369. doi:10.1016/J.ESWA.2019.03.019

Eyster, E., Rabin, M., & Vayanos, D. (2019). Financial markets where traders neglect the informational content of prices. *The Journal of Finance, 74*(1), 371-399. doi:10.1111/jofi.12729

Fama, E. F. (1991). Efficient capital markets: II. *The Journal of Finance, 46*(5), 1575-1617. doi:10.1111/J.1540-6261.1991.TB04636.X

Fama, E. F., & French, K. R. (2007). Disagreement, tastes, and asset prices. *Journal of Financial Economics, 83*(3), 667-689. doi:10.1016/j.jfineco.2006.01.003

Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of political Economy, 81*(3), 607-636. doi:10.1086/260061

Fang, L., & Peress, J. (2009). Media coverage and the cross-section of stock returns. *The Journal of Finance, 64*(5), 2023-2052. doi:10.1111/j.1540-6261.2009.01493.x

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research, 15*(1), 3133-3181.

Firth, M., Lin, C., & Wong, S. M. (2008). Leverage and investment under a state-owned bank lending environment: Evidence from China. *Journal of Corporate Finance, 14*(5), 642-653. doi:10.1016/j.jcorpfin.2008.08.002

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition*: Mcgraw-Hill Book Company.

Foroozani, A., & Ebrahimi, M. (2019). Anomalous information diffusion in social networks: Twitter and Digg. *Expert Systems With Applications*. doi:10.1016/J.ESWA.2019.05.047

French, K. R., Schwert, G. W., & Stambaugh, R. F. (1987). Expected stock returns and volatility. *Journal of Financial Economics, 19*(1), 3-29. doi:10.1016/0304-405X(87)90026-2

Gennaioli, N., Shleifer, A., & Vishny, R. (2012). Neglected risks, financial innovation, and financial fragility. *Journal of Financial Economics, 104*(3), 452-468. doi:10.1016/j.jfineco.2011.05.005

Giannini, R., Irvine, P., & Shu, T. (2019). The convergence and divergence of investors' opinions around earnings news: evidence from a social network. *Journal of Financial Markets, 42*, 94-120. doi:10.1016/j.finmar.2018.12.003

Giroud, X., & Mueller, H. M. (2010). Does corporate governance matter in competitive industries? *Journal of Financial Economics, 95*(3), 312-331. doi:10.1016/j.jfineco.2009.10.008

Giroud, X., & Mueller, H. M. (2011). Corporate governance, product market competition, and equity prices. *The Journal of Finance, 66*(2), 563-600. doi:10.1111/j.1540-6261.2010.01642.x

Goetzmann, W. N., Kim, D., & Shiller, R. J. (2017). Crash beliefs from investor surveys. *Working Paper 22143*. doi:10.3386/w22143

Golez, B., & Goyenko, R. (2019). Disagreement in the equity options market and stock returns. *SSRN Research Paper No. 3443241*. doi:10.2139/ssrn.3443241

Grčar, M., Cherepnalkoski, D., Mozetič, I., & Novak, P. K. (2017). Stance and influence of Twitter users regarding the Brexit referendum. *Computational social networks, 4*(1), 6. doi:10.1186/s40649-017-0042-6

Green, T. C., & Hwang, B.-H. (2009). Price-based return comovement. *Journal of Financial Economics, 93*(1), 37-50. doi:10.1016/j.jfineco.2008.09.002

Grullon, G., Kanatas, G., & Weston, J. P. (2004). Advertising, breadth of ownership, and liquidity. *The Review of Financial Studies, 17*(2), 439-461. doi:10.1093/rfs/hhg039

Han, B., & Yang, L. (2013). Social networks, information acquisition, and asset prices. *Management Science, 59*(6), 1444-1457. doi:10.1287/mnsc.1120.1678

Harris, M., & Raviv, A. (1993). Differences of opinion make a horse race. *The Review of Financial Studies, 6*(3), 473-506. doi:10.1093/rfs/5.3.473

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society, 47*(1), 153-161. doi:10.2307/1912352

Hendershott, T., Livdan, D., & Schürhoff, N. (2015). Are institutions informed about news? *Journal of Financial Economics, 117*(2), 249-287. doi:10.1016/j.jfineco.2015.03.007

Hermida, A. (2010). Twittering the news: The emergence of ambient journalism. *Journalism practice, 4*(3), 297-308. doi:10.1080/17512781003640703

Hillert, A., Jacobs, H., & Müller, S. (2014). Media makes momentum. *The Review of Financial Studies, 27*(12), 3467-3501. doi:10.1093/rfs/hhu061

Hirshleifer, D. (2001). Investor psychology and asset pricing. *The Journal of Finance, 56*(4), 1533-1597. doi:10.1111/0022-1082.00379

Hirshleifer, D. (2015). Behavioral finance. *Annual Review of Financial Economics, 7*(1), 133-159. doi:10.1146/annurev-financial-111914-043752

Hirshleifer, D. (Producer). (2019). Attention, psychological bias, and social interactions. *Presentation Slides, for Finance Theory Group Wharton Summer School (Long Version)*.

Hirshleifer, D. (2020). Presidential address: Social transmission bias in economics and finance. *The Journal of Finance, 75*(4), 1779-1831. doi:10.1111/jofi.12906

Hirshleifer, D., Lim, S. S., & Teoh, S. H. (2011). Limited investor attention and stock market misreactions to accounting information. *Review of Asset Pricing Studies, 1*(1), 35-73. doi:10.1093/rapstu/rar002

Hirshleifer, D., & Shumway, T. (2003). Good day sunshine: Stock returns and the weather. *The Journal of Finance, 58*(3), 1009-1032. doi:10.1111/1540-6261.00556

Hirshleifer, D., & Teoh, S. H. (2003). Limited attention, information disclosure, and financial reporting. *Journal of accounting and economics, 36*(1-3), 337-386. doi:10.1016/j.jacceco.2003.10.002

Hirshleifer, D., & Teoh, S. H. (2009). Thought and behavior contagion in capital markets. In *Handbook of Financial Markets: Dynamics and Evolution* (pp. 1-56): Elsevier.

Holden, C. W., & Jacobsen, S. (2014). Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions. *The Journal of Finance, 69*(4), 1747-1785. doi:10.1111/jofi.12127

Hong, H., Kubik, J. D., & Stein, J. C. (2004). Social interaction and stock-market participation. *The Journal of Finance, 59*(1), 137-163. doi:10.1111/j.1540-6261.2004.00629.x

Hong, H., & Stein, J. C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of Finance, 54*(6), 2143-2184. doi:10.1111/0022-1082.00184

Hong, H., & Stein, J. C. (2003). Differences of opinion, short-sales constraints, and market crashes. *The Review of Financial Studies, 16*(2), 487-525. doi:10.1093/rfs/hhg006

Hong, H., & Stein, J. C. (2007). Disagreement and the stock market. *Journal of economic perspectives, 21*(2), 109-128. doi:10.1257/jep.21.2.109

Howard, P. N., Duffy, A., Freelon, D., Hussain, M. M., Mari, W., & Maziad, M. (2011). Opening closed regimes: what was the role of social media during the Arab Spring?

Huang, D., Jiang, F., Tu, J., & Zhou, G. (2015). Investor sentiment aligned: A powerful predictor of stock returns. *The Review of Financial Studies, 28*(3), 791-837. doi:10.1093/rfs/hhu080

Huang, S., Huang, Y., & Lin, T.-C. (2019). Attention allocation and return co-movement: evidence from repeated natural experiments. *Journal of Financial Economics, 132*(2), 369-383. doi:10.1016/j.jfineco.2018.10.006

Huang, S., Maharjan, J., & Thakor, A. V. (2019). Disagreement-induced CEO turnover. *Journal of Financial Intermediation, 43*, 100819. doi:10.1016/j.jfi.2019.01.006

Huang, X., Nekrasov, A., & Teoh, S. H. (2018). Headline salience, managerial opportunism, and over-and underreactions to earnings. *The Accounting Review, 93*(6), 231-255. doi:10.2308/accr-52010

Hutton, A. P., Marcus, A. J., & Tehranian, H. (2009). Opaque financial reports, R2, and crash risk. *Journal of Financial Economics, 94*(1), 67-86. doi:10.1016/j.jfineco.2008.10.003

Hvidkjaer, S. (2006). A trade-based analysis of momentum. *The Review of Financial Studies, 19*(2), 457-491. doi:10.1093/rfs/hhj016

Hvidkjaer, S. (2008). Small trades and the cross-section of stock returns. *The Review of Financial Studies, 21*(3), 1123-1151. doi:10.1093/rfs/hhn049

Hwang, B. H. (2011). Country-specific sentiment and security prices. *Journal of Financial Economics, 100*(2), 382-401. doi:10.1016/j.jfineco.2010.10.020

Ivković, Z., & Weisbenner, S. (2005). Local does as local is: Information content of the geography of individual investors' common stock investments. *The Journal of Finance, 60*(1), 267-306. doi:10.1111/j.1540-6261.2005.00730.x

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance, 48*(1), 65-91. doi:10.1111/j.1540-6261.1993.tb04702.x

Jiang, J. X., Petroni, K. R., & Wang, I. Y. (2010). CFOs and CEOs: Who have the most influence on earnings management? *Journal of Financial Economics, 96*(3), 513-526. doi:10.1016/j.jfineco.2010.02.007

Jiang, L., Liu, J., & Yang, B. (2019). Communication and comovement: evidence from online stock forums. *Financial Management, 48*(3), 805-847. doi:10.1111/fima.12245

Jin, L., & Myers, S. C. (2006). R2 around the world: New theory and new tests. *Journal of Financial Economics, 79*(2), 257-292. doi:10.1016/j.jfineco.2004.11.003

Jones, J. J. (1991). Earnings management during import relief investigations. *Journal of Accounting Research, 29*(2), 193-228. doi:10.2307/2491047

Kahneman, D. (1973). *Attention and Effort* (Vol. 1063). Englewood Cliffs,New Jersey: Prentice-Hall, Inc.

Kai, L., & Prabhala, N. R. (2007). Self-selection models in corporate finance. In *Handbook of Empirical Corporate Finance* (pp. 37-86): Elsevier.

Kandel, E., & Pearson, N. D. (1995). Differential interpretation of public signals and trade in speculative markets. *Journal of political Economy, 103*(4), 831-872. doi:10.1086/262005

Karpoff, J. M. (1986). A theory of trading volume. *The Journal of Finance, 41*(5), 1069-1087. doi:10.1111/j.1540-6261.1986.tb02531.x

Khan, M., & Watts, R. L. (2009). Estimation and empirical properties of a firm-year measure of accounting conservatism. *Journal of accounting and economics, 48*(2-3), 132-150. doi:10.1016/j.jacceco.2009.08.002

Kim, J.-B., Li, Y., & Zhang, L. (2011a). Corporate tax avoidance and stock price crash risk: Firm-level analysis. *Journal of Financial Economics, 100*(3), 639-662. doi:10.1016/j.jfineco.2010.07.007

Kim, J.-B., Li, Y. H., & Zhang, L. D. (2011b). CFOs versus CEOs: Equity incentives and crashes. *Journal of Financial Economics, 101*(3), 713-730. doi:10.1016/j.jfineco.2011.03.013

Kim, J.-B., Lu, L. Y., & Yu, Y. (2019). Analyst coverage and expected crash risk: Evidence from exogenous changes in analyst coverage. *The Accounting Review, 94*(4), 345-364. doi:10.2308/accr-52280

Kim, J.-B., & Zhang, L. (2016). Accounting conservatism and stock price crash risk: Firm-level evidence. *Contemporary Accounting Research, 33*(1), 412-441. doi:10.1111/1911-3846.12112

Kim, O., & Verrecchia, R. E. (1991). Trading volume and price reactions to public announcements. *Journal of Accounting Research, 29*(2), 302-321. doi:10.2307/2491051

Kothari, S. P., Shu, S., & Wysocki, P. D. (2009). Do managers withhold bad news? *Journal of Accounting Research, 47*(1), 241-276. doi:10.1111/j.1475-679X.2008.00318.x

Kubick, T. R., & Lockhart, G. B. (2016). Proximity to the SEC and stock price crash risk. *Financial Management, 45*(2), 341-367. doi:10.1111/fima.12122

Kumar, A., & Lee, C. M. C. (2006). Retail investor sentiment and return comovements. *The Journal of Finance, 61*(5), 2451-2486. doi:DOI 10.1111/j.1540-6261.2006.01063.x

Kümpel, A. S., Karnowski, V., & Keyling, T. (2015). News sharing in social media: A review of current research on news sharing users, content, and networks. *Social media+ society, 1*(2), 14. doi:10.1177/2056305115610141

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). *What is Twitter, a social network or a news media?* Paper presented at the Proceedings of the 19th international conference on World wide web.

LaFond, R., & Watts, R. L. (2008). The information role of conservatism. *The Accounting Review, 83*(2), 447-478. doi:10.2308/accr.2008.83.2.447

Lee, C. M. C., & Radhakrishna, B. (2000). Inferring investor behavior: Evidence from TORQ data. *Journal of Financial Markets, 3*(2), 83-111.

Lee, C. M. C., & Ready, M. J. (1991). Inferring Trade Direction from Intraday Data. *The Journal of Finance, 46*(2), 733-746. doi:10.1111/j.1540-6261.1991.tb02683.x

Lee, D. S., & Mas, A. (2012). Long-run impacts of unions on firms: New evidence from financial markets, 1961–1999. *The Quarterly Journal of Economics, 127*(1), 333-378. doi:10.1093/qje/qjr058

Lee, E., Lee, J.-A., Moon, J. H., & Sung, Y. (2015). Pictures speak louder than words: Motivations for using Instagram. *Cyberpsychology, Behavior, and Social Networking, 18*(9), 552-556. doi:10.1089/cyber.2015.0157

Li, Q., Lourie, B., & Teoh, S. H. (2019). How salience of management guidance affects forecasting behavior: evidence from a quasi-natural experiment on estimize. *SSRN Research Paper No. 3423608*. doi:10.2139/ssrn.3423608

Li, S., & Zhan, X. T. (2019). Product market threats and stock crash risk. *Management Science, 65*(9), 4011-4031. doi:10.1287/mnsc.2017.3016

Liang, H. (2018). Broadcast versus viral spreading: The structure of diffusion cascades and selective sharing on social media. *Journal of Communication, 68*(3), 525-546. doi:10.1093/joc/jqy006

Lin, C., Ma, Y., & Xuan, Y. (2011). Ownership structure and financial constraints: evidence from a structural estimation. *Journal of Financial Economics, 102*(2), 416-431. doi:10.1016/j.jfineco.2011.06.001

Løwendahl, B., & Revang, Ø. (1998). Challenges to existing strategy theory in a postindustrial society. *Strategic Management Journal, 19*(8), 755-773. doi:10.1002/(SICI)1097-0266(199808)19:8<755::AID-SMJ968>3.0.CO;2-A

Luo, Zhang, J., & Duan, W. (2013). Social media and firm equity value. *Information Systems Research, 24*(1), 146-163. doi:10.1287/isre.1120.0462

Malmendier, U. (2018). Behavioral corporate finance. In B. D. Bernheim, S. Della Vigna, & D. Laibson (Eds.), *Handbook of Behavioral Economics: Applications and Foundations 1* (Vol. 1, pp. 277-379): Elsevier.

Markarian, G., & Parbonetti, A. (2007). Firm complexity and board of director composition. *Corporate Governance: An International Review, 15*(6), 1224-1243. doi:10.1111/j.1467-8683.2007.00643.x

Markowitz, H. (1952). The utility of wealth. *Journal of political Economy, 60*(2), 151-158. doi:10.1086/257177

Marks, H. (2011). *The Most Important Thing: Uncommon Sense for the Thoughtful Investor*: Columbia University Press.

Mathers, A. M., Wang, B., & Wang, X. S. (2017). Innovation and price informativeness. *Financial Management, 46*(2), 523-546. doi:10.1111/eufm.12134

Mazboudi, M., & Khalil, S. (2017). The attenuation effect of social media: Evidence from acquisitions by large firms. *Journal of Financial Stability, 28*, 115-124. doi:10.1016/j.jfs.2016.11.010

Miller, E. M. (1977). Risk, uncertainty, and divergence of opinion. *The Journal of Finance, 32*(4), 1151-1168. doi:10.1111/j.1540-6261.1977.tb03317.x

Modigliani, F., & Miller, M. H. (1958). The cost of capital, corporation finance and the theory of investment. *The American economic review, 48*(3), 261-297.

Morck, R., Yeung, B., & Yu, W. (2000). The information content of stock markets: Why do emerging markets have synchronous stock price movements? *Journal of Financial Economics, 58*(1-2), 215-260. doi:10.1016/S0304-405X(00)00071-4

Nekrasov, A., Teoh, S. H., & Wu, S. (2021). Visuals and attention to earnings news on Twitter. *Review of Accounting Studies, forthcoming*.

Newey, W. K., & West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review, 28*(3), 777-787.

Ngai, E. W., Tao, S. S., & Moon, K. K. (2015). Social media research: Theories, constructs, and conceptual frameworks. *International journal of information management, 35*(1), 33-44. doi:10.1016/j.ijinfomgt.2014.09.004

Nguyen, P.-A., & Kecskés, A. (2020). Do technology spillovers affect the corporate information environment? *Journal of Corporate Finance, 62*, 101581. doi:10.1016/j.jcorpfin.2020.101581

Ni, X., Peng, Q., Yin, S., & Zhang, T. (2020). Attention! Distracted institutional investors and stock price crash. *Journal of Corporate Finance, 64*, 101701. doi:10.1016/j.jcorpfin.2020.101701

Nisbett, R. E., & Ross, L. (1980). Human inference: Strategies and shortcomings of social judgment. *Journal of the American statistical association, 76*(374), 499-500. doi:10.2307/2287875

Odean, T. (1998). Volume, volatility, price, and profit when all traders are above average. *The Journal of Finance, 53*(6), 1887-1934. doi:10.1111/0022-1082.00078

Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems With Applications, 73*, 125-144. doi:10.1016/j.eswa.2016.12.036

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1992). Behavioral decision research: A constructive processing perspective. *Annual review of psychology, 43*(1), 87-131. doi:10.1146/ANNUREV.PS.43.020192.000511

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The Adaptive Decision Maker*: Cambridge University Press.

Peng, L., & Xiong, W. (2006). Investor attention, overconfidence and category learning. *Journal of Financial Economics, 80*(3), 563-602. doi:10.1016/j.jfineco.2005.05.003

Peress, J. (2014). The media and the diffusion of information in financial markets: Evidence from newspaper strikes. *The Journal of Finance, 69*(5), 2007-2043. doi:10.1111/jofi.12179

Petersen, M. A. (2009). Estimating standard errors in finance panel data sets: Comparing approaches. *The Review of Financial Studies, 22*(1), 435-480. doi:10.1093/rfs/hhn053

Pindyck, R. S. (1984). Risk, inflation, and the stock market. *The American economic review, 74*(3), 335-351.

Pindyck, R. S., & Rotemberg, J. J. (1993). The comovement of stock prices. *The Quarterly Journal of Economics, 108*(4), 1073-1104. doi:10.2307/2118460

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

Porta, R. L., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. W. (1998). Law and finance. *Journal of political Economy, 106*(6), 1113-1155. doi:10.1086/250042

Pound, J., & Zeckhauser, R. (1990). Clearly heard on the street - The effect of takeover rumors on stock-prices. *Journal of Business, 63*(3), 291-308. doi:10.1086/296508

Rainie, L., Brenner, J., & Purcell, K. (2012). Photos and videos as social currency online.

Ranco, G., Aleksovski, D., Caldarelli, G., Grcar, M., & Mozetic, I. (2015). The Effects of Twitter Sentiment on Stock Price Returns. *Plos One, 10*(9), e0138441. doi:10.1371/journal.pone.0138441

Renault, T. (2017). Intraday online investor sentiment and return patterns in the US stock market. *Journal of Banking & Finance, 84*, 25-40. doi:10.1016/j.jbankfin.2017.07.002

Roll, R. (1988). R2. *The Journal of Finance, 43*(3), 541-566. doi:10.1111/j.1540-6261.1988.tb04591.x

Romer, D. (1993). Rational asset-price movements without news. *The American economic review*, 1112-1130.

Roychowdhury, S., & Sletten, E. (2012). Voluntary disclosure incentives and earnings informativeness. *The Accounting Review, 87*(5), 1679-1708. doi:10.2308/accr-50189

Rubinstein, A. (1993). On price recognition and computational complexity in a monopolistic model. *Journal of political Economy, 101*(3), 473-484. doi:10.1086/261883

Saunders, E. M. (1993). Stock prices and Wall Street weather. *The American economic review, 83*(5), 1337-1345.

Shanthikumar, D. M. (2012). Consecutive earnings surprises: Small and large trader reactions. *The Accounting Review, 87*(5), 1709-1736. doi:10.2308/ACCR-50188

Sheng, J. (2019). *Macro news, micro news, and stock prices.* Paper presented at the 44th Annual Meeting of the European Finance Association, Mannheim, Germany.

Shiller, R. J. (1980). Do stock prices move too much to be justified by subsequent changes in dividends? *Working Paper*. doi:10.3386/w0456

Shiller, R. J. (1981). The use of volatility measures in assessing market-efficiency. *The Journal of Finance, 36*(2), 291-304. doi:10.1111/j.1540-6261.1981.tb00441.x

Shiller, R. J. (1989). Comovements in stock prices and comovements in dividends. *The Journal of Finance, 44*(3), 719-729. doi:10.1111/j.1540-6261.1989.tb04387.x

Shiller, R. J. (1992). *Market Volatility*. London: MIT Press.

Shiller, R. J. (1995). Conversation, information, and herd behavior. *The American economic review, 85*(2), 181-185.

Shleifer, A. (2000). *Inefficient Markets: An Introduction to Behavioural Finance*. Oxford: Oxford University Press.

Shleifer, A., & Summers, L. H. (1990). The noise trader approach to finance. *Journal of economic perspectives, 4*(2), 19-33. doi:10.1257/jep.4.2.19

Sila, V., Gonzalez, A., & Hagendorff, J. (2017). Independent director reputation incentives and stock price informativeness. *Journal of Corporate Finance, 47*, 219-235. doi:10.1016/j.jcorpfin.2017.09.018

Skinner, D. J., & Sloan, R. G. (2002). Earnings surprises, growth expectations, and stock returns or don't let an earnings torpedo sink your portfolio. *Review of Accounting Studies, 7*(2), 289-312. doi:10.1023/A:1020294523516

Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management, 20*(5), 926-957. doi:10.2139/ssrn.1702854

Sprenger, T. O., & Welpe, I. M. (2011). Tweets and peers: Defining industry groups and strategic peers based on investor perceptions of stocks on Twitter. *Algorithmic Finance, 1*(1), 57-76. doi:10.3233/AF-2011-006

Staiger, D., & Stock, J. H. (1994). Instrumental variables regression with weak instruments. *Working Paper*. doi:10.3386/t0151

Stambaugh, R. F., Yu, J., & Yuan, Y. (2012). The short of it: Investor sentiment and anomalies. *Journal of Financial Economics, 104*(2), 288-302. doi:10.1016/j.jfineco.2011.12.001

Stock, J., & Yogo, M. (2005). Asymptotic distributions of instrumental variables statistics with many instruments.

Surowiecki, J. (2005). *The Wisdom of Crowds*. Canada: Anchor.

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter networks. *Social networks, 34*(1), 73-81. doi:10.1016/j.socnet.2011.05.006

Tumarkin, R., & Whitelaw, R. F. (2001). News or noise? Internet postings and stock prices. *Financial Analysts Journal, 57*(3), 41-51. doi:10.2469/faj.v57.n3.2449

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207-232. doi:10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty - Heuristics and biases. *Science, 185*(4157), 1124-1131. doi:10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453-458. doi:10.1007/978-1-4613-2391-4_2

Varian, H. R. (1985). Divergence of opinion in complete markets: a note. *The Journal of Finance, 40*(1), 309-317. doi:10.1111/j.1540-6261.1985.tb04951.x

Varian, H. R. (1989). Differences of opinion in financial markets. In C. Stone (Ed.), *Financial Risk: Theory, Evidence and Implications* (pp. 3-37): Springer.

Vega, C. (2006). Stock price reaction to public and private information. *Journal of Financial Economics, 82*(1), 103-133. doi:10.1016/j.jfineco.2005.07.011

Veldkamp, L. L. (2006). Information markets and the comovement of asset prices. *The review of Economic studies, 73*(3), 823-845. doi:10.1111/j.1467-937X.2006.00397.x

Verrecchia, R. E. (1983). Discretionary disclosure. *Journal of accounting and economics, 5*, 179-194. doi:10.1016/0165-4101(83)90011-3

Watts, R. L. (2003). Conservatism in accounting part I: Explanations and implications. *Accounting horizons, 17*(3), 207-221. doi:10.2308/acch.2003.17.3.207

Wood, A., & Pasquier, V. (2018). The power of social media as a labour campaigning tool: Lessons from Our Walmart and the Fight for 15. (10).

Wysocki, P. D. (1998). Cheap talk on the web: The determinants of postings on stock message boards. *University of Michigan Business School Working Paper*(98025).

Yu, F. F. (2008). Analyst coverage and earnings management. *Journal of Financial Economics, 88*(2), 245-271. doi:10.1016/j.jfineco.2007.05.008

Yu, J. (2011). Disagreement and return predictability of stock portfolios. *Journal of Financial Economics, 99*(1), 162-183. doi:10.1016/j.jfineco.2010.08.004

Yu, Y., Duan, W. J., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems, 55*(4), 919-926. doi:10.1016/j.dss.2012.12.028

Yuan, Y. (2008). *Attention and trading*. University of Pennsylvania. Philadelphia.

Yuan, Y. (2015). Market-wide attention, trading, and stock returns. *Journal of Financial Economics, 116*(3), 548-564. doi:10.1016/J.JFINECO.2015.03.006

Zhao, D., & Rosson, M. B. (2009). *How and why people Twitter: The role that micro-blogging plays in informal communication at work.* Paper presented at the Proceedings of the ACM 2009 international conference on supporting group work.

Ethical Approval Letter

The University Of Sheffield.

Downloaded: 08/07/2021
Approved: 04/11/2019

Mustabsar Awais
Registration number: 160130881
Management School
Programme: PhD

Dear Mustabsar

PROJECT TITLE: Essays in Corporate Finance: Investigating the Role of Social Media in Financial Markets
APPLICATION: Reference Number 019102

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 04/11/2019 the above-named project was **approved** on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 019102 (form submission date: 23/10/2019); (expected project end date: 01/10/2020).

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since written approval will be required.

Your responsibilities in delivering this research project are set out at the end of this letter.

Yours sincerely

Sophie May
Ethics Administrator
Management School

Please note the following responsibilities of the researcher in delivering the research project:

- The project must abide by the University's Research Ethics Policy: https://www.sheffield.ac.uk/rs/ethicsandintegrity/ethicspolicy/approval-procedure
- The project must abide by the University's Good Research & Innovation Practices Policy: https://www.sheffield.ac.uk/polopoly_fs/1.671066!/file/GRIPPolicy.pdf
- The researcher must inform their supervisor (in the case of a student) or Ethics Administrator (in the case of a member of staff) of any significant changes to the project or the approved documentation.
- The researcher must comply with the requirements of the law and relevant guidelines relating to security and confidentiality of personal data.
- The researcher is responsible for effectively managing the data collected both during and after the end of the project in line with best practice, and any relevant legislative, regulatory or contractual requirements.