8-2019

Collaborative Speculation and Overvaluation: Evidence from Social Media

Adam Barrett Booker
University of Arkansas, Fayetteville

Collaborative Speculation and Overvaluation: Evidence from Social Media

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Business Administration

by

Adam Booker
University of Arkansas
Bachelor of Science in Electrical Engineering, 2008
University of Arkansas
Master of Information Systems, 2014
University of Arkansas
Master of Accountancy, 2015

August 2019
University of Arkansas

Vernon Richardson, PhD
Dissertation Director

Gary Peters, PhD
Committee Member

Mike Crawley, PhD
Committee Member

Asher Curtis, PhD
Ex-officio Member

Abstract

I use data from StockTwits and Twitter to provide evidence that investor attention on social media in the period before earnings is related to short-term overvaluation, consistent with bullish investors herding around common information. In the 2 to 60 days after earnings, returns for companies in the highest quintile of pre-earnings announcement investor attention are 4.2 percent lower than those of companies in the lowest quintile. I find evidence that the negative post-earnings drift result found in this study is related to investors waiting until after earnings are announced to enact costly arbitrage strategies. I further examine intra- and inter-network herding and find evidence that social media influences investors beyond the population of active users. This study contributes to prior literature on herding, social media, and speculation and arbitrage.

Acknowledgements

I would like to thank Vernon Richardson and Asher Curtis for their continuous help and guidance. I would like to thank Mike Crawley and Gary Peters for support and help with this dissertation and for serving on my committee. I would like to thank Pierce Crosby for access to StockTwits data.

Dedication

I dedicate this dissertation to Benjamin and Skylor Booker.

Table of Contents

1 Introduction

This study examines the association between investor attention on StockTwits and Twitter in the period preceding earnings announcements and negative equity market returns in the period after earnings are announced. This is important because social media has changed the way that investors access and process information, lowering the costs of information acquisition but exposing investors to information with questionable accuracy or focus. Investor attention on social media, as opposed to investor attention on search engines or investor attention as proxied by volume, is inherently collaborative because users can interact and view one another's posts (Bartov et al. 2018). Prior literature has found that there is online, public information that can partially preempt earnings (Bagnoli et al. 1999, Bartov et al. 2018). I contribute to this literature by providing evidence consistent with attention-constrained investors using social media to research and identify stocks to buy, which results in increasing short-term overvaluations as more investors pay attention and act on a common set of information.

The internet has changed the financial information ecosystem and reduced information and investing frictions for retail investors. Starting with message boards and more recently continuing with social media, the internet has made it easier for investors to work together to process information. Ideally, these connectivity platforms help improve the information available to and understanding of information among investors, especially retail investors. My results suggest, however, that investor attention on social media as proxied by number of active StockTwits users in the pre-earnings-announcement period (-10 to -2 days relative to earnings) is related to temporary overvaluation. These results are consistent with prior literature on herding through social learning (Fudenberg and Kreps 1995) and imitation (Banerjee 1992, Bikhchandani et al. 1992, Shiller 1995, Hirshleifer and Teoh 2003).

While this study does incorporate a set of posts from Twitter, the main focus is on data from StockTwits, a financial social media platform. StockTwits, per their Google search headline "Share Ideas & Learn from Passionate Investors & Traders" is based on the premise that users will influence each other.[1] StockTwits is like other social media sites, such as Twitter or Facebook, in that users are able to directly communicate with each other opening the possibility of users influencing collective opinions. Due to this collaborative environment, active participation on StockTwits can be contrasted to other internet-based proxies for attention like Google search volume where investors are not necessarily provided with and influenced by insights from other investors. I find evidence that could be interpreted as active users' posts influence investors that are not actively participating as well.

I use social media data to examine investor attention and sentiment in the period before earnings. Consistent with the results in Bartov et al. (2018), I find that the average sentiment from the group of users discussing stocks in the period before earnings is positively associated with earnings period returns, whereas investor attention, defined as the number of active users in the pre-earnings period, is negatively associated with both the earnings period returns and post-earnings period returns, This indicates the crowd on social media is often discussing relevant earnings information, but that attention is related to investor herding around stale information and overshooting expectations, causing short-term overvaluation. I also test for and find evidence that overconfidence (Oskamp 1965, Khaneman and Tversky 1974) related to the number of different topics covered in posts in the pre-earnings-announcement period dominates the

[1] In conversations with StockTwits users at their annual meeting, Stocktoberfest, many told me that they learned how to trade and are making money with knowledge they gained from the website, consistent with the StockTwits website title: "StockTwits © - Share Ideas & Learn from Passionate Investors & Traders."

association between social media attention and overvaluation, indicating the influence of social media goes beyond users that are actively participating.

As noted in Barber and Odean (2007):

> How can we measure the extent to which a stock grabs investors' attention? A direct measure would be to go back in time and, each day, question the hundreds of thousands of investors in our datasets as to which stocks they thought about that day. Since we cannot measure the daily attention paid to stocks directly, we do so indirectly (p. 787).[2]

Prior literature has relied on proxies such as trading volume, media attention and extreme returns to proxy for events that capture investors' attention. These proxies, however, are indirect measures of investor attention. Abnormal trading volume is indicative of the final decision made by investors but does not consider investors that paid attention but did not act. Media attention does not consider how investors react to news. Extreme returns may be the result of attention as well as material information. I argue that investor attention on social media is a more direct measure of investor attention because it 1) identifies investors' response to news, and 2) includes investors that pay attention but don't trade. This is important because it allows for measurement of the information qualities that lead to short-term speculative overvaluation. I do not find evidence that when the sentiment of the crowd is negative there is undervaluation and attribute this to short-selling constraints.

I expect that StockTwits caters primarily to retail investors who focus on picking stocks to buy rather than short sell. This could be in part due to the increased costs for short-selling (D'Avolio 2002). Consistent with overvaluation due to a segment of investors using social media to research stocks to buy, I find that post-earnings (pre-earnings) period returns are significantly

[2] See also Klibanoff et al. (1998): "The underlying problem facing financial economists is that neither fundamentals nor other possible determinants of investor behavior, such as 'investor sentiment,' are observable."

negatively (positively) associated with investor attention when sentiment is positive but are not associated with investor attention when sentiment is negative. I support this by providing cross sectional evidence that the attention-overvaluation association is only significant when both traditional media and social media sentiment are positive, consistent with investors being influenced by a bullish feedback loop. In other cross-sectional tests, I find that the association between overvaluation and investor attention in the pre-earnings-announcement period is more pronounced in samples with lower institutional ownership, which provides evidence that that arbitragers postpone their strategies when limited by short-selling constraints.

Returns in the 60-day period after earnings announcements are negatively associated with investor attention, regardless of earnings surprise. Miller (1977) explains this anomaly with a model in which price is a function of collective opinion. In pre-earnings announcement periods where information is correct but possibly stale or noisy, price formation is delayed as bearish, short-selling constrained investors are kept out of the market. In this model, based on the equilibrium point where supply meets demand, price increases as the bullish group of investors herds around stale information and causes price to overshoot its fundamental value. Chen et al. (2001) and Kelly and Tetlock (2013) provide expanded theoretical models for the discussion in Miller (1977).

I make three main contributions with this study. First, I add to the literature on herding by examining the market impact of investor attention to publicly viewable information. The ease with which information can be disseminated on social media (Blankespoor et al. 2014) makes it an ideal platform for low-cost information acquisition by investors. This low-cost barrier makes it an especially viable source of information for less sophisticated investors who may have knowledge limitations and cost barriers not shared by their institutional counterparts. In Figure 1, I find that the relationship between attention on social media and negative post-earnings return drift is near-monotonic across quintiles of investor attention.

Second, I elaborate on the results in Bartov et al. (2018) which finds that sentiment on social media is predictive of earnings surprise and returns. This finding shows that valuable information is being shared on social media, which makes is it an ideal platform for information acquisition. StockTwits feeds at the time of this writing are incorporated in platforms including: Bloomberg, Thomson Reuters, Interactive Brokers, Fidelity, Charles Schwab, Trading Technologies, and eSignal. Social media has been shown in the literature and demonstrated by its use in trading platforms to be a viable source of financial information. However, it seems the market is not efficiently incorporating the level of attention given to stocks, as evidenced by the negative post-earnings drift found in this paper. This should be of interest to investors and regulators, since as SEC Chairman Jay Clayton said: “serving and protecting Main Street investors is my main priority at the SEC” (SEC 2018).

Third, I contribute to the literature on speculative trading and arbitrage. Brunnermier and Abreu (2002, 2003) develop a model in which investors do not know when mispricing will be corrected. Short-horizon investors predict not only fundamentals, but also the behavior of other investors. Social media offers an ideal setting to test this. With the assumption that a segment of the investing population is short-selling constrained by fees, search problems or offsetting capital (D’Avolio 2002) investors using social media to research stocks to buy in advance of earnings, in line with Miller (1977), cause a temporary price increase. When earnings are announced, these investors are no longer focused on earnings, which allows price to return to fundamentals.

The remainder of the paper is organized as follows. Section 2 presents a review of the literature and hypothesis development. Section 3 describes the research design. Section 4 details the data used and descriptive statistics. Section 5 presents the empirical results, and I conclude in Section 6.

2 Literature Review and Hypothesis Development

2.1 StockTwits as a Financial Information Aggregator

Hayek (1945) gave one of the first economics-based arguments for decentralized aggregation of information. At the time, economists were involved in conversations about whether a centralized economy or a distributed economy was the most efficient. Hayek argues that decentralized decisions are better able to incorporate idiosyncratic information. Hayek (1945) is supported by studies that investigate countries that split up after World War 2. For example, the unsuccessful centralized economy in East Germany versus the more successful decentralized economy in West Germany. Social media extends this concept and can be contrasted with centralized or traditional media. Today, social media posts are often featured in traditional media stories and media outlets and media articles are often disseminated on social media. Media decentralization has gotten to a point where it is difficult to distinguish traditional media from social media.

StockTwits has become a valuable source of information about stocks as it allows for an efficient method of aggregating information about stocks from a heterogeneous group of users. The founders of StockTwits invented cashtags, a way for investors to qualify words as tickers by prepending a dollar sign to the ticker. This seemingly simple idea allows for the efficient aggregation of company-specific information and therefore more efficient communication among investors. The cashtag is now a widely-accepted way to qualify ticker symbols and is used on StockTwits as well as Twitter, with Twitter adopting a few years after the success on StockTwits. Finding company information about Agilent using ticker symbol “A” is extremely noisy. It is computationally difficult to discern the much more common article “a” from the ticker for Agilent. However, searching for “$A” on Google will lead to StockTwits and other results about Agilent.

The posts on social media leverage the power of decentralization and aggregation. Investors can go to social media to see what people are saying and to get an idea about current news. The commentaries on social media about current news help add color for investors without access to insiders, staff or private trading platforms.

2.2 Herding and Independence

Hirshleifer and Teoh (2003), in their review of information cascades and herding, give a thorough development of the behavioral components that lead to correlated actions among individuals. They define herding as a convergence in behavior and cascading as ignoring private information. They include direct communication and observational influence as possible sources of herding. Cascading can be thought of as a unit autoregressive process where each action is determined only by the last action. For example, if a group of investors only used the trade at time t_{-1} as the basis the next trade at time t_0 then each subsequent action would be the same for all t. Banerjee (1992) and Bikhchandani et. al. (1992) propose cascading as the mechanism responsible for causing herding. Shiller (1995) applies these concepts to a social setting in which herding is based on convergence to group norms through social interaction, as would possibly occur on social media.

Models of herding have a commonality in that they all rely on a lack of independence between individuals in a group. On StockTwits as more investors share information and have discussions in the pre-earnings period, these investors are potentially exerting increasing influence on one another through their social interactions. StockTwits is often a top result in search engines[3] and is used in many trading platforms, so even investors that are not active on StockTwits are also potentially exposed to the information from these social interactions. I assume that investor attention on StockTwits is correlated with the overall level of market attention, and investor attention on StockTwits is likely to influence less sophisticated investors, who would be more likely to ignore their private information as in models of herding. When these investors are focusing on stocks to buy and there are market-wide arbitrage constraints, this would lead to overvaluation as new bullish investors enter the equities market and bearish investors are kept out of the market.

Curtis et al. (2016) find evidence that contemporaneous earnings-period investor attention is related to more rapid price discovery around earnings announcements, which is consistent social media aiding the alignment investors' opinions. I find that social media sentiment in the pre-earnings period is positively associated with earnings-period returns, which provides evidence that social media users are providing actionable information. However, as more investors acquire information from social media and contribute to social media, there is potentially a self-reinforcing feedback mechanism in which investors act on bullish sentiment, price subsequently increases, speculators are validated, and the security gets more attention in the short-term only to decrease in price after the earnings event. In addition to online interactions, StockTwits users interact in physical social situations. Figure 1 presents the StockTwits meetups in North America. This figure shows that 13,437 people where potentially meeting up in 2017 to discuss how to make money in the stock market. These face-to-face meetings may further degrade the independence between users.

[3] StockTwits is a top result for about 2/3 of the sample used in this study and in the sample of all tickers in 2014 and 2015 available on CRSP.

2.3 Social Media and the Wisdom of Crowds

Bartov et al. (2018) examine opinion formation on social media using constructs from the wisdom of crowds. The genesis of the term wisdom of crowds can be traced to Dr. Francis Galton's 1907 study titled "Vox Populi" in which individuals' guesses about the weight of a dressed ox at a local fair were used as data to provide empirical evidence on the wisdom of crowds. He found that when the guesses of individuals were averaged, the collective guess was near perfect. This finding is congruent with the Central Limit Theorem which states that when a group of observations is independent and identically distributed, the observations converge to a normal distribution as the number of observations increases. In the case of peoples' guesses, ideally, this distribution would be around the actual value. The difference between the setting at the fair and the setting on social media is that social media users use the platform explicitly for communication whereas the attendees at the fair did not have information about one another's guesses.

The Wisdom of Crowds by James Surowiecki (2004) gives 4 necessary conditions for wise crowds: diversity of opinion, independence, decentralization, and aggregation.[4] Surowiecki (2004) offers several instances when crowds are not wise because they violate some element of the wisdom of crowds. In the case of social media there is diversity of opinion, decentralization and aggregation, but users can influence each other and those that view their public interactions and thereby potentially violate the independence condition. In a setting with investors focused on picking stocks to buy, this would lead to overvaluation as investors herd around information

[4] Surowiecki (2004) has 566 citations from papers that are indexed on Google scholar and include the terms "earnings" or "accounting."

indicating bullish outcomes. Following Miller (1977), overpricing increases as bullish investors establish new positions in a stock while bearish investors are kept out of the market.

2.4 Speculative Trading and Arbitrage

Brunnermier and Abreu (2002, 2003) provide theoretical models in which investors face a synchronization risk when arbitraging mispricing. If arbitrageurs act immediately they incur holding costs that can make their strategy prohibitive. So, instead they try to time their strategies so as to minimize holding costs. When a sufficient number of arbitrageurs synchronize their timing, price begins to return to fundamental value. An alternate explanation for the results in this paper is that post-earnings liquidity trading causes price to return to fundamental value (Bernard and Thomas 1989, 1990). In Bernard and Thomas (1989, 1990) transaction costs make it prohibitive for arbitrageurs to act on their knowledge of mispricing and price slowly drifts toward the fundamental price after earnings are announced.

Grinblatt and Keloharju (2000) find evidence that retail investors in Finland are net buyers of momentum stocks with weak future performance and that institutional investors are net buyers of momentum stocks with strong future performance. In contrast, Boehmer et al. (2016) find that stocks with the highest positive (negative) order imbalance[5] in the prior week have the highest positive (negative) abnormal returns in the next 20 days. Lawrence et al. (2018) find evidence of a causal link between advertisements on Yahoo! and contemporaneous positive abnormal returns when companies beat analyst earnings expectations. This study complements Lawrence et al (2018) by providing evidence that investor attention in speculative periods leads

[5] Defined as dollar value of shares bought less the dollar value of shares sold scaled by the average daily trading volume over the prior year.

to overvaluation. Short selling constraints offer a reason for why there is an association between attention and overvaluation and not undervaluation.

Regulation T provides that brokers that lend shares to short sellers must have a “bona fide” cash deposit to offset the lent shares. Firms that lend shares also charge interest on the shares further increasing the short selling premium. Short selling agreements can vary among brokers, but they generally include a fee discount for the cash collateral. The interest rates in the proprietary set used in Reed (2002) range from the special rate of 7.6 percent to the regular rate of 5.8 percent. D’Avolio (2002), also using a proprietary dataset, shows that the special rate can be as high as 79 percent. Almazan et al. (2004) note that “73.3 percent of the 679 funds that filed Form N-SAR in 1994 reported that their investment policies formally restricted them from selling short” (p. 9). These significant frictions offer a plausible explanation for why investors herding around bullish social media information can be difficult to arbitrage.

In addition to extra costs, there can be search problems for investors that want to sell short. Prior literature has found that the search costs for finding a lender decrease with the level of institutional ownership (Almazan et al. 2004, D’Avolio 2002). D’Avolio (2002) finds that the special rate increases with attention on Yahoo! Message Boards. As barriers to arbitrage increase, I expect that mispricing will be more sensitive to investor attention on social media.

Models of short selling constraints in prior literature assume prices are a weighted average of beliefs (supply and demand) from a heterogeneous set of investors (Miller 1977). Shares sold short are borrowed, in effect increasing the supply of shares in the market and theoretically shifting the intersection between supply and demand to the right, lowering price. Short sales constraints inflate prices by forcing bearish investors out of the market (Lintner 1969, Miller 1977, Figlewski 1981, Jarrow 1981). The short-term overpricing effects of short sales constraints increase as the level of disagreement between bullish and bearish investors increases.

I draw upon the models of short-selling constrained investors in Chen et al. (2001) and Kelly and Tetlock (2013), which build on models in Miller (1977). The simple theoretical model in Kelly and Tetlock (2013) examines the investing population using three sets of investors: A) investors that are 100 percent rational without short-selling constraints that always pay attention, B) investors that are less than 100 percent rational without short-selling constraints that always pay attention, C) investors that are less than 100 percent rational with short-selling constraints that are not 100 percent attentive. The market clearing price (or price at which the demand from investors in {A, B, C} equals supply) includes a mispricing term that is proportional to sentiment and attention. The mispricing term in Kelly and Tetlock (2013) is a function of sentiment, attention, short-selling constraints, disagreement and risk tolerance. When there are short-selling constraints, overpricing increases with sentiment and attention. This leads to my hypothesis, stated in the alternative:

H1: Investor attention on social media in the pre-earnings-announcement period is positively related to overvaluation.

3 Empirical Design

I start the empirical design section with a description of the textual analysis methods incorporated in this study. Textual analysis, while not directly applicable to the main variable of interest, investor attention on social media, is a component of the experimental design and thus necessary to explain for later exposition of the models employed.

3.1 Sentiment Measurement

I use a supervised machine learning method known as the Paragraph Vector method (Mikolov and Le, 2014) to classify the level of bullishness of unclassified posts in the StockTwits and Twitter datasets used in this study. This method utilizes a neural network that incorporates the position of words used in documents. This is in contrast to bag of words methods that assume independence between words. The Paragraph Vector method is ideal for StockTwits data because of the approximately 8 million user-classified posts in the dataset that facilitate supervised learning. In untabulated results, I find that Paragraph Vector-based classification is more accurate than Naïve Bayes classification. The Paragraph Vector classification used in this paper is 64% accurate in the full set of posts. The Naïve Bayes classification using the same training data is 55% accurate. Both methods are more accurate with smaller datasets. Training with millions of user-classified records introduces noise to the extent that users say the same thing with different explicit opinion qualifiers. For example, if User A posts "I am excited about $A earnings" (user-classified as bullish) and User B posts "I am excited about $A earnings" (user-classified as bearish), these posts introduce noise in the classification model.

To quantify user sentiment, each unclassified post is given a value that indicates the probability of the post being either positive (bullish) or negative (bearish). If a post is user-classified as bullish, it is given a sentiment value of 1. If a post is user-classified as bearish, it is given a sentiment value of -1. The paragraph vectors from the set of user-classified bullish and bearish posts are used to calculate the parameters of a logistic regression model. I use these parameters with the Paragraph Vectors of unclassified posts to calculate the probabilities of unclassified posts being either bullish or bearish. I use the average sentiment from each user in the pre-earnings period. *SM Sent* is the sum of these averages and is calculated as follows:

$$SM\ Sent_t = \frac{1}{\#Users} \sum_i Average\ Sentiment_i$$

Where *i* is the index for each post from a given user. *SM Sent* is the by-user average sentiment in the pre-announcement period *t*.

3.2 Using LDA to Measure User-Generated Information

Shiller (2017) uses the term “narrative economics” to describe how narratives influence economies and provides a discussion of how people think in terms of narratives. Thorsrud (2018) uses Latent Dirichlet Allocation (LDA) learned topics as a proxy for narratives in news stories and shows that LDA topics and sentiment in media articles can be used to predict economic fluctuations. I use LDA to classify the topics that social media users are discussing around earnings. This unsupervised learning algorithm uncovers latent topics using the underlying frequencies of the words used in documents. Although the label of the topics used in this study does not influence the results, I manually add labels to the topics to help make the divisions more salient. The topics include technical, earnings speculation, past earnings info, attention to news, recommendations, SEC filings and earnings news.

LDA assumes documents are combinations of topics and topics are based on word frequency distributions. The LDA algorithm uses sampling based on known priors (word frequencies and number of topics) to train model parameters that maximize the tradeoff between the precision of document topic distributions and precision of topic word frequency distributions. Chang et al. (2009) finds that LDA classifies topics as a person would.

I train the LDA model used in this study on the set of all posts in the -10 to -2 days relative to earnings announcements so as to best capture topics discussed in the period before earnings. Prior to training the model, I convert all cashtags from their specific form to a generic form (e.g. $A to $cashtag), convert all mentions (e.g. @bob to @mention), convert all links (e.g. http://www.yahoo.com to |link), convert numbers (e.g. 123 to #number) and lemmatize all words (e.g. stopping to stop). I train the model with the cleaned set of posts. I then use the trained model to derive the most representative topic for each post. For each quarterly announcement I construct a variable *Ln(#Topics)* which equals the natural log of the number of topics discussed in the pre-earnings period.

I vary the number of topics between 20 and 80 and find similar results with these variations. I use 40 topic categories in this study. Topics with written-in names are provided in Appendix B. The categories include words that indicate topics including: current news, past news, user speculation, technical charts, and earnings news.

In line with prior literature that has provided evidence that retail investors are more likely to buy rather than short a stock (for example, Barber and Odean 2007), and with literature on information and overconfidence. Oskamp (1965) was one of the first studies to show that people become more confident as the amount of information they use in a decision increases, even if new information is not material. In his experimental study, subjects were given information about a case study. As the subjects were given more (but not relevant) info, their confidence about being correct rose from 33 to 53% while their accuracy remained consistently below 30%. They describe a similar study on the "representativeness" bias: confidence in new, but worthless information without regard to prior probabilities. Subjects in the representativeness study were told that 2/3 of a population were engineers and 1/3 lawyers. When the subjects were given additional but worthless information their guesses were around 50%, whereas without that additional information, the subjects' guesses were closer to the prior probabilities.

3.3 Empirical Models

3.3.1 Main Model

The main research question concerns whether there is an association between attention on social media and short-term overvaluation. I therefore begin the empirical analysis with an examination of the association between the number of active users in the pre-earnings-announcement period (-10 to -2 days relative to the earnings announcement) and abnormal returns. I examine the attention-returns association in three timespans: -10 to -2 days relative to earnings, -1 to 1 days around earnings, and 2 to 60 days[6] after earnings are announced. The pre-earnings timespan captures the contemporaneous association between social media attention and returns. The earnings period timespan captures the preemptive nature of the association between attention and returns. The post earnings timespan is of interest for answering the main research question, whether social media attention leads to short-term overvaluation. This model assumes that price adjusts to the arrival of fundamental information in the earnings period, but does not completely resolve overvaluation from the shift in the investing population. If the overvaluation were resolved in the earnings period then there should be no drift in the post earnings period. I used the following model, Model 1 to test for an association between overvaluation and attention:

$$AR_{i,j,t} = \alpha + \boldsymbol{\beta_1 Ln(\#Users)_{i,j,[-10,-2]}} + \sum_k \beta_k Control_{k,i,j} + e_{i,j,t}$$

[6] I follow Bernard and Thomas (1989), which finds that most drift occurs in the 60 days after earnings are announced, I examine abnormal returns in the 2 to 60 days after earnings are announced to identify return reversals.

AR are the market-adjusted abnormal returns in period *t* relative to the quarterly earnings announcement *i* for firm *j*. The main variable of interest in this model is $Log(\#Users)_{-10,-2}$. I expect that β_1 will be positive in the period before earnings are announced and negative in the period after earnings are announced, consistent with short-term overvaluation and subsequent reversal. I test H1 in the period after earnings are announced, when investors are no longer have incentive to speculate on earnings.

I support the association between investors using social media and the search for stocks to buy and overvaluation with an examination of investor attention and order imbalance. I use 2 different measures of order imbalance: 1) overall order imbalance with buys and sells classified using the algorithm in Chakrabarty et al. (2007), and 2) retail order imbalance as classified by the algorithm in Boehmer et al. (2017). The Chakrabarty et al. (2007) algorithm extends the Lee and Ready (1993) algorithm to better account for trades that occur inside the bid and ask quotes, which controls for misclassification of stocks that are shorted. Boehmer et al. (2017) relies on retail orders being filled through a wholesale market which can be uncovered with fraction of cent trades on the FINRA Trade Reporting Facility. I examine the association between investor attention and buying stocks using Model 2, shown below:

$$OIMB_{i,j,t} = \alpha + \boldsymbol{\beta_1 Ln}(\#\boldsymbol{Users})_{\boldsymbol{i,j},[-\mathbf{10},-\mathbf{2}]} + \sum_k \beta_k Control_{k,i,j} + e_{i,j,t}$$

OIMB is the order imbalance in period *i* relative to the quarterly earnings announcement at time *t* for firm *j*. I expect that β_1 will be positive in the period before earnings are announced for retail order imbalance. In support of H1, I expect that β_1 will be positive for overall and retail order imbalance in the period before earnings. I expect that overall order imbalance will be negative in the period after earnings are announced.

I examine the association between social media attention and synchronization risk for investors that do not currently own the overvalued stocks using short selling data from the NASADQ PSX exchange. Because there are many additional costs such as having offsetting capital, margin payments when price fluctuates, and opportunity costs (D'Avolio 2002), it is beneficial for short sellers to wait until after earnings are announced to sell short. I use Model 3, described below, to test this:

$$Short\ Interest_{i,j,t} = \alpha + \boldsymbol{\beta_1 Ln(\#Users)_{i,j,[-10,-2]}} + \sum_k \beta_k Control_{k,i,j} + e_{i,j,t}$$

Where *Short Interest* is the number of shared sold short scaled by shares outstanding in period *i* relative to the quarterly earnings announcement at time *t* for firm *j*. If arbitrageurs are timing their strategies, I expect that the coefficient on β_1 will be more negative in the period after earnings are announced.

3.3.2 Intra- or Inter-Network Herding?

If social media users are herding around information that is pushed to their accounts, then I expect that attention from users with the most visibility (proxied for by their number of followers) will have a larger negative association with post-earnings returns than the total level of social media attention. However, since StockTwits is often a first-page Google search result and is incorporated in several trading platforms, I might not find evidence of within-network herding. For each year in the StockTwits dataset, I rank StockTwits users by their number of followers (people that subscribe to their posts) and label the top 1000 users. I use this set of top users to measure the effects of intra-network herding. Model 4 is shown below:

$$AR_{i,j,t} = \alpha + \boldsymbol{\beta_1 Ln}(\#\boldsymbol{Users})_{\mathbf{-10,-2}} + \boldsymbol{\beta_2 Ln}(\#\boldsymbol{Top\ Followed\ Users})_{\mathbf{-10,-2}} + \sum_k \beta_k Control_{k,i,j} + e_{i,j,t}$$

AR are the market-adjusted abnormal returns in period *i* relative to the quarterly earnings announcement at time *t* for firm *j*. If only investors within the StockTwits platform are buying, then I expect the coefficient from the users with the most followers, β_2, will be greater than the coefficient on $\beta_1$in the post-earnings period.

If instead, investors that don't participate, but use the information on StockTwits, are driving the short-term overvaluation then I expect that these users will be more likely to buy a stock based on the number of topics covered in the posts that they browse, consistent with prior behavioral findings in Oskamp (1965) and Khaneman and Tversky (1974). I use Model 5, shown below to examine inter-network herding resulting from overconfidence:

$$AR_{i,j,t} = \alpha + \boldsymbol{\beta_1 Ln}(\#\boldsymbol{Users})_{\mathbf{-10,-2}} + \boldsymbol{\beta_2 Ln}(\#\boldsymbol{Topics})_{\mathbf{-10,-2}} + \sum_k \beta_k Control_{k,i,j} + e_{i,j,t}$$

If investors that come across StockTwits (this includes investors that are members of StockTwits as well as investors that incorporate StockTwits chatter in their decisions) are convinced to buy because of a greater number of topics, then I expect that β_2 will be negative and significant in the post-earnings period. If investors outside of the active StockTwits community become overconfident as a result of the information shared on the site, then β_2 will be less than β_1 in the post-earnings period.

3.4 Control Variables

I add controls for firm characteristics and sources of public information other than social media. I add *Market Cap*, the natural logarithm of the number of shares outstanding multiplied by the share price at the end of the fiscal quarter to control for lower information content in the earnings news for larger firms (Ataise 1985). I add *Ln(#Analysts)*, the natural logarithm of the analyst following in the I/B/E/S Summary database, to control for the demand for information from sophisticated investors (Bhushan 1989).[7] I add *Market to Book*, calculated using Compustat quarterly market and book values, to control for value or earnings yield (Penman et al. 1996, Ball et al. 2017). I add *Institutional%* to control for the level of investor sophistication (Potter 1992) and short-selling constraints (Miller 1977). I add a Q4 indicator variable to control for information differences in the 4th fiscal quarter (Das and Shroff 2002). I add standardized unexpected earnings (*SUE*) to control for the impact of new information about cash flows.

I add additional controls for media attention and sentiment, StockTwits sentiment, and attention to the SEC's EDGAR website. I proxy for media attention with the natural log of the number of news article observations in the RavenPack database (*Ln(#News Stories)*). For each observation in the RavenPack database, I center the RavenPack Composite Sentiment Score (CSS) around zero and rescale the range to [-1, 1]. I take the average of the scaled sentiment in all news stories to proxy for the news sentiment (*News Sent*). I include *SM Sent* to control for the average sentiment of users on StockTwits. Finally, I include *Ln(Retail EDGAR IP)* to control for investors' demand for fundamental information. I calculate *Ln(Retail EDGAR IP)* using the EDGAR IP logs and according to the algorithm in Drake et al. (2015).

4 Data and Descriptive Statistics

The data sources and variables are listed in Appendix A. Social media data are from StockTwits and Twitter. StockTwits data were granted by StockTwits management. StockTwits

---

[7] The firms used in this study have at least on analyst following them during the given firm-quarter. In untabulated tests, social media activity from firms with no analyst following has a weaker (not statistically significant) correlation with equity market returns.

data consists of the complete set of posts in the years 2014 and 2015. In addition to the post itself, each record includes, among other things, attributes of the user, the time of the post, user sentiment, and cashtags used. StockTwits has no limitation on stocks that can be discussed, so the final sample consists of all companies that have active users participating in the -10 to -2 days relative to earnings that can be merged with the other data in the study. StockTwits posts at the time of this writing are incorporated in Bloomberg, Thomson Reuters, Interactive Brokers, Fidelity, Charles Schwab, Trading Technologies, eSignal, and other online trading platforms. An example post is provided in Figure 5. StockTwits is a private company the earns revenue from advertising and events.

Twitter data were collected using the streaming API in the period from August to December 2015. This dataset includes 99.9% of the posts for which a cashtag in the Standard and Poor's 1500 was used. A very small portion of tweets were lost due to rate limiting by the Twitter API.

I use CRSP to calculate abnormal returns and market capitalization. I obtain analyst following for each firm from the I/B/E/S summary database. I calculate the percentage holdings by institutional investors using the Thomson Reuters Institutional Holdings database, which is derived from the 13-F filings of institutional investors. I use the Trades and Quotes (TAQ) millisecond intraday data to construct the measures of order imbalance used in this study. I use Compustat to construct the measure of market to book. I use RavenPack[8] to construct measures of news attention and sentiment. Finally, I use EDGAR IP Log files to construct measures of attention to fundamental information.

[8] "RavenPack analyzes unstructured content from thousands of publications to extract information on named entities and financially relevant events in the public eye" (http://www.whartonwrds.com/datasets/ravenpack/).

Table 1 presents the descriptive statistics for the main results. The average abnormal returns in the -10 to -2 days relative to earnings (the pre-earnings-announcement period) as well as in the 2 to 60 days after earnings are slightly negative. The main results in this study are robust to the inclusion of firm-fixed effects. The mean earnings period returns are near zero. The median quarterly earnings announcement had 5 unique active StockTwits users in the pre-earnings-announcement period. The median of 5 users is similar to the median seen in Table 7 of Bartov et al. (2018). The median quarterly earnings announcement had 72 retail EDGAR IP accesses and 164 different news observations in RavenPack. The median market to book ratio in the sample is 2.574, which is consistent with social media users following glamor stocks. The median analyst following is 9. The maximum institutional ownership in the sample is greater than 1, which Lewellen (2009) attributes to 13F data only including long positions (that is, shares held and lent out for short-selling are included). The portion of the sample with over 100 percent institutional ownership is also consistent with Lewellen (2009). The mean standardized unexpected earnings (*SUE*) is positive. The *Q4* indicator is 1 about 25 percent of the time, consistent with a near-balanced panel of quarterly announcements.

Table 2 presents the Fama-French 48 industry classifications for the sample used to construct the main results. Business Services has the most of observations relative to other categories. Business Services has more subcategories than many of the other Fama-French classification and encompasses computers and technology. Blankespoor et al. (2014) limit their sample to technology firms because these firms have better representation on social media. This is consistent with my sample. Agriculture has the fewest observations, also consistent with social media users being more likely to be interested in technology-like firms.

5 Results

5.1 Simulation and Actual

Panel A of Figure 2 presents the results of a simulation based on the theoretical short-selling constrained equilibrium price in Kelly and Tetlock (2013) and Panel B presents the actual market-adjusted abnormal returns contemporaneous with quintiles of StockTwits sentiment and attention. The comparison between the two panels assumes that the stocks in Panel B were at their fundamental price at the beginning of the pre-earnings-announcement period and no new fundamental information has been released in the pre-earnings period. With this assumption, the contemporaneous abnormal returns associated with investor attention and sentiment on StockTwits align with the simulation of the model of misvaluation in Kelly and Tetlock (2013).

5.2 Main Results

The central research question in this study relates to whether attention on social media in the period before earnings is associated with overvaluation. Stated formally, is attention on social media in the speculative period before earnings associated with short-term overvaluation? Obviously, each post is from someone conveying information that may be or may not be relevant, but it is an empirical question whether or not the number of users participating is related to attention to stale information. In Figure 3, I provide compelling evidence that attention on social media is related to short-term overvaluation. I use 3-Factor plus momentum-adjusted returns in graphs to control for firm characteristics that are covary with other firms and cause noise. When separated by quintile of attention on StockTwits, I find that post-earnings drift is monotonically decreasing in the 3 quintiles with the most attention, with an abnormal returns of -4.2 %, -1.8% and -0.5%. The 2 quintiles with the last attention have near-zero post-earnings drift.

Table 3 presents the results of Model 1. In Model 1, the dependent variable is market-adjusted abnormal returns. The table is divided into three periods: the pre-earnings period (-10 to -2 days before earnings), the earnings period (-1 to 1 days around earnings), and the post-earnings period (2 to 60 days after earnings). The results show that StockTwits $Ln(\#Users)_{-10,-2}$ is positively related to returns in the period before earnings (coefficient=0.004, t-statistic=3.24) are announced. This finding is consistent with short-selling constrained investors' stock purchasing actions being associated with social media activity. The coefficient on $Ln(\#Users)_{-10,-2}$ during the earnings period is negative and significant in the earnings period (coefficient=-0.002, t-statistic=-2.02), consistent with prior literature that finds there is online information that preempts earnings news (Bagnoli et al. 1999, Bartov et al. 2018). The coefficient on $Ln(\#Users)_{-10,-2}$ is also negative and significant in the post-earnings period (coefficient=-0.01, t-statistic=-4.69), providing additional support for H1.

Among the noteworthy controls in Table 3, in the post-earnings period the coefficients on each of the other proxies for attention are insignificant. $Ln(\#News\ Stories)_{-10,-2}$ controls for the cumulative news coverage and is negative in the pre-earnings-announcement period as well as during the earnings period and is not significant in the period after earnings are announced. $Ln(\#News\ Stories)_{-10,-2}$ does not capture the investors' reaction to news stories. I also include a control for the number of retail views (by IP address) of EDGAR information to proxy for investor attention to fundamental information. This proxy does capture interest from fundamental investors, however, the coefficient in the post-earnings period is insignificant. Other investors do not see the interpretations of EDGAR views and therefore this null result for the coefficient on $Ln(\#Retail\ EDGAR\ IP)_{-10,-2}$ is expected to be zero.

Both the average sentiment from news outlets as well as the average sentiment from StockTwits users in the pre-earnings announcement period is positively related to earnings period abnormal returns. The finding that news sentiment in 2014 and 2015 is related to earnings period returns is in contrast to findings in Bartov et al. (2018) which does not find this relationship in 2009-2012. This speaks to the increasing overlap in content between traditional and social media.

In Table 4, I use Twitter posts from the Standard and Poor's 1500 in the period from August 2015 and December 2015 to show that the relationship between attention and overvaluation is consistent across social media platforms. The dependent variable is market-adjusted abnormal returns in both models. The coefficient on $Ln(\#Users)_{-10,-2}$ is similar in magnitude in both models, -0.007 on StockTwits and -0.008 on Twitter. The $Ln(\#Users)_{-10,-2}$ coefficient is more significant in the model using Twitter data, which could be attributed to Twitter having better coverage of the investing population. This intuitively makes sense as Twitter has a much larger user base than StockTwits and therefore is a less noisy approximation of the aggregate level of investor attention. The results also indicate that Twitter sentiment is a better predictor of earnings period returns. The coefficient on Twitter sentiment is 0.108 with a t-statistic of 3.31 whereas the coefficient on StockTwits sentiment is 0.024 with a t-statistic of 0.72. These results suggest that social media attention in general before earnings is related to overvaluation, even when controlling for contemporaneous pre-earnings returns.

In Table 5, I present the results of Model 2. I use two measures of order imbalance as my dependent variables: retail order imbalance and overall order imbalance. $Ln(\#Users)_{-10,-2}$ is positively related to retail order imbalance in both the pre-earnings-announcement period (coefficient=0.021, t-statistic=9.60) and the post-earnings-announcement period (coefficient=0.020, t-statistic=9.66). However, $Ln(\#Users)_{-10,-2}$ is only significantly related to overall order imbalance in the pre-earnings-announcement period. This finding is consistent with social media attention capturing the contemporaneous exuberance of the market, but also with the findings in Lee (1992) in which retail investors are likely to buy stocks without considering earnings surprise because these events draw the attention of retail investors. $Ln(\#Users)_{-10,-2}$ is not significantly related to overall earnings period or post-earnings period order imbalance. The lack of a significant relationship suggests that current owners are not arbitraging short-term overvalued stocks in this sample. $Ln(\#News\ Stories)_{-10,-2}$ is positively related to retail order imbalance and negatively related to overall order imbalance in the pre-earnings-announcement period. This is consistent with prior literature that finds evidence that institutional investors provide liquidity to retail investors focusing on glamour stocks (Barber and Odean 2007).

In Table 6, I present the results of Model 3 in which the relationship between short selling and social media activity around earnings announcements is examined. Abreu and Brunnermier (2002) provide a model in which arbitrageurs delay until there is lower synchronization risk. If arbitrageurs wait until after earnings are announced to enact their strategies, I expect that the relationship between attention on social media and short selling will be greater after earnings are announced. In Table 9, I use daily short selling data from the NASDAQ PSX exchange to provide evidence that this is the case. The relationship between $Ln(\#Users)_{-10,-2}$ and short selling is positive and significant in all periods. Using the Z-statistic to test for differences between models (Clogg et al. 1995), the coefficient in the period after earnings is significantly greater than in the period before earnings are announced (p-value = 0.09), providing additional evidence that could be could be construed as arbitrageurs delaying their strategies until after earnings are announced. These results suggest that synchronization risk is part of the reason for the negative post-earnings drift associated with social media attention. That is, the current owners of the stocks do not seem to be selling after earnings are announced, but a population of arbitrageurs that don't currently own the stocks in this study appear to be waiting until after earnings to sell short.

Table 7 presents cross-sectional results based on quintile sorts. The dependent variable in all models is abnormal returns in the 2 to 60 days relative to earnings. In panels with quintile sorts, quintile 1 encompasses the portion of the sample with the lowest levels of the sort variable and quintile 5 encompasses the portion of the sample with the highest levels of the sort variable. Panel A presents results based on sorts of institutional ownership. Institutional ownership has been used as a proxy for shares available to be sold short (D'Avolio 2002, Asquith et al. 2005). The coefficient on $Ln(\#Users)_{-10,-2}$ is negative and significant in quintiles 1-3 (coefficients=-0.014, -0.011, -0.013 for quintiles 1, 2, 3) and not significant in quintiles 4 and 5 (coefficients=-0.003,-0.001 for quintiles 4, 5), which have the greatest institutional ownership. These results provide evidence consistent with synchronization risk when the resolution of short-term overvaluation is delayed when it is more costly.

I examine disagreement as proxied by analyst EPS forecast dispersion in Panel B and find evidence that overvaluation increases with uncertainty among analysts, consistent with models based on Miller (1977) and with Boehme et al. (2006), which finds that overvaluation is related to the dispersion of opinions among investors. I present the results of sorts based on analyst following in Panel C and do not find significant variation across these samples. Panel D presents the results based on sorts by the level of dispersion of opinions from social media users in the -2 to -10 days relative to earnings. $Ln(\#Users)_{-10,-2}$ is negative and significant in quintile 5 (coefficient=-0.018, t-statistic=-3.03) and insignificant in quintile 1 (coefficient=0.003, t-statistic=0.47). This result is consistent with the analyst dispersion result in Panel B.

In Panel E, I present results of a cross sectional analysis based on media sentiment and social media sentiment. I find that investor attention in the pre-announcement period is only related to future negative returns when both media and social media sentiment are positive (coefficient=-0.013, t-statistic=-4.96). This provides evidence that investors are using multiple streams of information and are more likely to buy stocks when information in these different channels is aligned. This finding supports the feedback loop explanation for herding in Fudenberg and Kreps (1995).

In Table 8, I separate the sample by positive and negative social media sentiment to examine whether there is a difference between investor attention when sentiment is generally negative as opposed to when sentiment is generally positive. Consistent with investor attention being related to picking stocks to buy and social media users influencing each other, I find that abnormal returns in the post-earnings-announcement period are negatively related to investor attention when sentiment is positive (coefficient=-0.013, t-statistic=-4.59) and are not related when sentiment is negative (coefficient=-0.005, t-statistic=--1.57). Sentiment predicts returns only in the set of announcement in which investors were positive.

I further examine social media attention when sentiment is positive sentiment in Figure 4. Figure 4.1 provides the full-period (-10 to 60 days around earnings) graph of average 3 factor plus momentum-adjusted returns for companies that beat analyst expectations for the highest and lowest quintiles of social media attention. The stocks that had the least attention prior to earnings, have almost no drift after earnings are announced. The set of stocks that had the highest levels of social media attention have distinctly negative drift after earnings are announced. The high attention set of companies also seems to be preempting earnings, but only to later reverse even beyond the price increase from the earnings surprise. Figure 4.2 provides the results for companies that missed analyst expectations. This set of companies also has negative drift after earnings are announced.

5.3 Intra- or Inter-Network Herding?

In Tables 9 and 10, I present the results of Models 4 and 5 and provide evidence that passive users[9] are contributing to the herding effects seen in this study. In Table 9, I include attention from the most visible users in the network as well as attention from the all active users. StockTwits users with the most followers will have their posts viewed by more other users within the StockTwits network.[10] If active StockTwits users are driving this result then I expect attention from top users have a greater effect on herding since users within the platform are more likely to see these top users' posts. In Table 10, I examine the relationship between the number or topics covered across all posts and overvaluation. If investors that are either active or passively participating in StockTwits are driving the results in this study then I expect the number of topics covered will have a greater impact on overvaluation as investors are more likely to become overconfident in this setting (Khaneman and Tversky 1974).

Table 9 provides the results of Model 4 and provides evidence that the most visible users on StockTwits are not driving the reversal effect. The coefficient on $\mathit{Ln(\#Users)}_{-10,-2}$ is positive and significant (coefficient=-0.007, t-statistic=-2.22) whereas the coefficient on *Ln(#Top Users)* is not significant (coefficient=0.003, t-statistic=0.68). The top-followed users are the most visible

[9] Passive users in this context are not active in the StockTwits community in the period before earnings. StockTwits is a first-page Google search result for two-thirds of the cashtags used in this study, so it is possible that StockTwits chatter is being utilized by investors that are not members of the site. StockTwits feeds are also utilized in Bloomberg, Thomson Reuters, Interactive Brokers, Fidelity, Charles Schwab, Trading Technologies, and eSignal.
[10] If a user follows another user the follower will see the posts from the followee.

to users within the network. By design their posts are pushed to the greatest number of other users. These within-network highly-visible users have the same visibility to investors that come across StockTwits as part of their information search. This result suggests that intra-network herding is not driving the result found in this study.

Table 10 presents the results of Model 5 and provides evidence that investors that use StockTwits (active or passive) are more likely to buy stocks when more topics are covered in the pre-earnings-announcement period, consistent with investors becoming increasingly overconfident as more information is presented in the pre-earnings-announcement period. The number of topics discussed by users in the pre-earnings-announcement period, $Ln(\#Topics)_{-10,-2}$ (coefficient=-0.013, t-statistic=-3.32) dominates $Ln(\#Users)_{-10,-2}$ (coefficient=-0.001, t-statistic=-0.39) in the post-earnings period. At the extreme, if all 40 topics are covered, this would lead to a 5 percent negative abnormal return in the period after earnings are announced. This provides evidence that inter-network herding among the set of investors that utilize social media as an information source is driving the negative drift associated with social media attention. The results in Table 9 in combination with the results in Table 10, are consistent with passive StockTwits users contributing to the overvaluation effect seen in this study. This provides evidence that herding on social media spills over to investors that don't actively participate.

5.4 The Chicken or the Egg

So far this study has left as an open question whether social media users are following the market or whether the market is following social media users. That is, generally speaking, are social media users discussing events that have already occurred or are they discussing relevant speculative information? I examine this question using Grainger causality tests for the complete sample timespan (2014 and 2015) and in a panel of the 60 days after earnings are announced.

In Table 11, I present the results of this analysis. I find that change in social media attention in the full timespan Granger-causes returns (chi-squared=36.461) and returns Granger-cause change in social media attention (chi-squared=11.921), indicating a general cointegrated relationship between social media attention and returns. However, in the panel set constructed on the 60 days after earnings are announced, I find that returns Granger-cause change in social media attention (chi-squared=5.483), but change in social media attention does not Granger-cause returns. These results could be interpreted to suggest that StockTwits users are more focused on speculation than post mortem analysis of earnings information.

6 Conclusion

This study uses data from StockTwits and Twitter to examine whether investor attention on social media is related to short-term overvaluation in the speculative period before earnings are announced. I find evidence that social media users (as well as news sources in RavenPack) are generally correct about upcoming earnings announcements, but that attention on social media is related to short-term overvaluation, providing evidence that social media users, active or passive, are paying attention to and acting on stale information. This finding is of interest to investors, regulators and academics.

This study finds that attention from StockTwits users is related to contemporaneous overall order imbalance, but not related to overall order imbalance after earnings are announced. This is consistent with current owners of stocks not selling their overvalued stocks, which provides evidence that part of the negative drift result found in this study is not attributable to synchronization risk from current owners. I also find that the relationship between attention from StockTwits users has a greater relationship with short-selling after earnings are announced, indicating that arbitrageurs are timing their strategies. Together these results suggest that the negative drift found in this study is related arbitrage from short sellers.

This study finds evidence that the influence of social media spills over to users that are not actively participating in conversations. This is important because it demonstrates that social media users are influencing investors that they may not be aware of. This spill-over of discussions to trading platforms or investors that come across social media is an interesting area for future study that may help add sophistication and understanding to the impacts of this relatively new information dissemination platform.

This study contributes to literature on heading through social learning, to literature on social media and to literature on speculation and arbitrage. This study does not intend to imply that social media should be regulated. There were speculative bubbles long before social media. For example, the Tulip bubble that collapsed in 1637 (Shiller 2005), over 200 years before the invention of the telegraph.

7. References

Almazan, Andres, Keith C. Brown, Murray Carlson, and David A. Chapman. "Why constrain your mutual fund manager?." *Journal of Financial Economics 73*, no. 2 (2004): 289-321.

Asquith, Paul, Parag A. Pathak, and Jay R. Ritter. "Short interest, institutional ownership, and stock returns." *Journal of Financial Economics* 78, no. 2 (2005): 243-276.

Atiase, Rowland Kwame. "Predisclosure information, firm capitalization, and security price behavior around earnings announcements." *Journal of Accounting Research* (1985): 21-36.

Bagnoli, Mark, Messod D. Beneish, and Susan G. Watts. "Whisper forecasts of quarterly earnings per share." *Journal of Accounting and Economics 28*, no. 1 (1999): 27-50.

Ball, Ray and Gerakos, Joseph J. and Linnainmaa, Juhani T. and Nikolaev , Valeri V., Earnings, Retained Earnings, and Book-to-Market in the Cross Section of Expected Returns (September 25, 2017). *Chicago Booth Research Paper No. 17-03*. Available at SSRN: https://ssrn.com/abstract=2924798 or http://dx.doi.org/10.2139/ssrn.2924798

Banerjee, Abhijit V. "A simple model of herd behavior." *The Quarterly Journal of Economics 107*, no. 3 (1992): 797-817.

Barber, Brad M., and Terrance Odean. "All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors." *The review of financial studies* 21, no. 2 (2007): 785-818.

Bartov, Eli, Lucile Faurel, and Partha S. Mohanram. "Can Twitter help predict firm-level earnings and stock returns?." *The Accounting Review 93*, no. 3 (2018): 25-57.

Bernard, Victor L., and Jacob K. Thomas. "Post-earnings-announcement drift: delayed price response or risk premium?." *Journal of Accounting Research* (1989): 1-36.

Bernard, Victor L., and Jacob K. Thomas. "Evidence that stock prices do not fully reflect the implications of current earnings for future earnings." *Journal of Accounting and Economics* 13, no. 4 (1990): 305-340.

Bhushan, Ravi. "Firm characteristics and analyst following." *Journal of Accounting and Economics* 11, no. 2 (1989): 255-274.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. "A theory of fads, fashion, custom, and cultural change as informational cascades." *Journal of Political Economy 100*, no. 5 (1992): 992-1026.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3, no. Jan (2003): 993-1022.

Boehme, Rodney D., Bartley R. Danielsen, and Sorin M. Sorescu. "Short-sale constraints, differences of opinion, and overvaluation." *Journal of Financial and Quantitative Analysis 41, no. 2* (2006): 455-487.

Boehmer, Ekkehart, Charles Jones, and Xiaoyan Zhang. "Tracking retail investor activity." (2017).

Abreu, Dilip, and Markus K. Brunnermeier. "Synchronization risk and delayed arbitrage." *Journal of Financial Economics* 66, no. 2-3 (2002): 341-360.

Abreu, Dilip, and Markus K. Brunnermeier. "Bubbles and crashes." *Econometrica* 71, no. 1 (2003): 173-204.

Chakrabarty, Bidisha, Bingguang Li, Vanthuan Nguyen, and Robert A. Van Ness. "Trade classification algorithms for electronic communications network trades." *Journal of Banking & Finance* 31, no. 12 (2007): 3806-3821.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. "Reading tea leaves: How humans interpret topic models." *Advances in neural information processing systems*, pp. 288-296. 2009.

Clogg, Clifford C., Eva Petkova, and Adamantios Haritou. "Statistical methods for comparing regression coefficients between models." *American Journal of Sociology* 100, no. 5 (1995): 1261-1293.

Curtis, Asher, Vernon J. Richardson, and Roy Schmardebeck. "Investor attention and the pricing of earnings news." (2016), in *Handbook of Sentiment Analysis in Finance*, eds.

Das, S., and P. K. Shroff. “Fourth Quarter Reversals in Earnings Changes and Earnings Management.” *Working paper*, University of Minnesota, 2002.

D’avolio, Gene. "The market for borrowing stock." *Journal of Financial Economics* 66, no. 2-3 (2002): 271-306.

Drake, Michael S., Darren T. Roulstone, and Jacob R. Thornock. "Investor information demand: Evidence from Google searches around earnings announcements." *Journal of Accounting Research 50*, no. 4 (2012): 1001-1040.

Drake, Michael S., Darren T. Roulstone, and Jacob R. Thornock. "The determinants and consequences of information acquisition via EDGAR." *Contemporary Accounting Research 32*, no. 3 (2015): 1128-1161.

Engelberg, Joseph E., and Christopher A. Parsons. "The causal impact of media in financial markets." *The Journal of Finance 66*, no. 1 (2011): 67-97.

Figlewski, Stephen. "The informational effects of restrictions on short sales: Some empirical evidence." *Journal of Financial and Quantitative Analysis 16*, no. 4 (1981): 463-476.

Fudenberg, Drew, and David M. Kreps. "Learning in extensive-form games I. Self-confirming equilibria." *Games and Economic Behavior 8*, no. 1 (1995): 20-55.

Galton, Francis. "Vox populi (The wisdom of crowds)." *Nature 75*, no. 7 (1907): 450-451.

Grinblatt, Mark, and Matti Keloharju. "The investment behavior and performance of various investor types: a study of Finland's unique data set." *Journal of Financial Economics 55*, no. 1 (2000): 43-67.

Hayek, Friedrich August. "The use of knowledge in society." *The American Economic Review 35*, no. 4 (1945): 519-530.

Hirshleifer, David, and Siew Hong Teoh. "Herd behaviour and cascading in capital markets: A review and synthesis." *European Financial Management 9*, no. 1 (2003): 25-66.

Hirshleifer, David, Sonya Seongyeon Lim, and Siew Hong Teoh. "Driven to distraction: Extraneous events and underreaction to earnings news." *The Journal of Finance* 64, no. 5 (2009): 2289-2325.

Jarrow, Robert A., and George S. Oldfield. "Forward contracts and futures contracts." *Journal of Financial Economics 9, no. 4* (1981): 373-382.

Kahneman, Daniel and Amos Tversky. "Judgment under uncertainty: Heuristics and biases." *Science 185*, no. 4157 (1974): 1124-1131.

Keynes, John Maynard. "The general theory of employment." *The Quarterly Journal of Economics 51*, no. 2 (1937): 209-223.

Kelley, Eric K., and Paul C. Tetlock. "Retail short selling and stock prices." *Working paper, no. 3* (2013): 801-834. Available at https://www0.gsb.columbia.edu/mygsb/faculty/research/pubfiles/6123/Kelley%20Tetlock%20-%20Nov%202013%20-%20Retail%20Short%20Selling%20and%20Stock%20Prices.pdf

Klibanoff, Peter, Owen Lamont, and Thierry A. Wizman. "Investor reaction to salient news in closed-end country funds." *The Journal of Finance 53*, no. 2 (1998): 673-699.

Lewellen, Jonathan, 2011, “Institutional Investors and the Limits of Arbitrage", *Journal of Financial Economic*s, 102, 62-80. http://faculty.tuck.dartmouth.edu/images/uploads/faculty/jonathan-lewellen/Institutions.pdf

Lawrence, Alastair and Ryans, James and Sun, Estelle and Laptev, Nikolay, Earnings Announcement Promotions: A Yahoo Finance Field Experiment (February 6, 2018). Available at SSRN: https://ssrn.com/abstract=2940223 or http://dx.doi.org/10.2139/ssrn.2940223

Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." *In International Conference on Machine Learning*, pp. 1188-1196. 2014.

Lee, Charles MC. "Earnings news and small traders: An intraday analysis." *Journal of Accounting and Economics* 15, no. 2-3 (1992): 265-302.

Lee, Charles MC, and Mark J. Ready. "Inferring trade direction from intraday data." *The Journal of Finance 46*, no. 2 (1991): 733-746.

Lewellen, Jonathan, 2011, "Institutional Investors and the Limits of Arbitrage", *Journal of Financial Economics*, 102, 62-80. http://faculty.tuck.dartmouth.edu/images/uploads/faculty/jonathan-lewellen/Institutions.pdf

Lintner, John. "The aggregation of investor's diverse judgments and preferences in purely competitive security markets." *Journal of Financial and Quantitative Analysis 4, no. 4* (1969): 347-400.

Miller, Edward M. "Risk, uncertainty, and divergence of opinion." *The Journal of Finance 32*, no. 4 (1977): 1151-1168.

Oskamp, Stuart (1965). "Overconfidence in case-study judgments" (PDF). *Journal of Consulting Psychology*. 29 (3): 261–265. doi:10.1037/h0022125.

Peng, Lin, and Wei Xiong. "Investor attention, overconfidence and category learning." *Journal of Financial Economics* 80, no. 3 (2006): 563-602.

Penman, Stephen H. "The articulation of price-earnings ratios and market-to-book ratios and the evaluation of growth." *Journal of Accounting Research* (1996): 235-259.

Potter, Gordon. "Accounting earnings announcements, institutional investor concentration, and common stock returns." *Journal of Accounting Research* (1992): 146-155.

Reed, Adam V. "Costly short-selling and stock price adjustment to earnings announcements." PhD diss., *University of Pennsylvania*, 2002.

Shiller, Robert J. "Conversation, Information, and Herd Behavior." *The American Economic Review 85*, no. 2 (1995): 181-85. http://www.jstor.org/stable/2117915.

Shiller, Robert J. (2005), *Irrational Exuberance* (2nd ed.), Princeton: Princeton University Press, ISBN 0-691-12335-7

Shiller, Robert J. "Narrative economics." *American Economic Review 107, no. 4* (2017): 967-1004.

Surowiecki, James. "The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business." *Economies, Societies and Nations 296* (2004).

SEC 2018, https://www.sec.gov/fact-sheets/sec-protects-retail-investors-markets

Tetlock, Paul C. "Giving content to investor sentiment: The role of media in the stock market." *The Journal of Finance 62*, no. 3 (2007): 1139-1168.

Thorsrud, Leif Anders. "Words are the New Numbers: A Newsy Coincident Index of the Business Cycle." *Journal of Business & Economic Statistics* just-accepted (2018): 1-35

Weng, Jianshu, Ee-Peng Lim, Jing Jiang, and Qi He. "Twitterrank: finding topic-sensitive influential twitterers." In *Proceedings of the third ACM international conference on Web search and data mining*, pp. 261-270. ACM, 2010.

8. Appendices

Appendix A
Variable descriptions

Variable	Description	Database
Abnormal Returns	The CRSP return (ret) less the value weighted return (vwretd)	CRSP
Order Imbalance (OIMB)	The value of stock bought less the value of stock sold divided by the total value of stock bought and sold: (buy-sell)/(buy+sell). Buys and sells are classified according to the algorithm in Chakrabarty et al. (2007)	TAQ
Retail Order Imbalance (Retail OIMB)	The value of stock bought by retail traders less the value of stock sold by retail traders. Trades classified according to the algorithm in Boehmer et al. (2016)	TAQ
Short Interest	The number of shared sold short scaled by the number of shares outstanding	
Ln(#Users)$_{-10,-2}$	The natural Ln of the number of unique active users in the -10 to -2 days before earnings are announced	StockTwits
Ln(#Top Users)	The natural log of the number of users in the set of 1000 users with the most followers in a given year	StockTwits
Ln(#Topics)	The entropy of the topics that StockTwits users discuss in a given time period.	StockTwits
Ln(#News Stories)	The natural Ln of the number of news stories	RavenPack
Ln(#Retail EDGAR IP)	The natural Ln of the number of EDGAR IP accesses from retail users using the algorithm from Drake et al. (2015)	EDGAR IP Logs

Appendix A (Cont.)

Variable	Description	Database
Ln(#Analysts)	The natural log of the number of analysts following a firm during the given quarter	I/B/E/S Summary
Market Cap	The shares outstanding multiplied by stock price the day before earnings are announced	CRSP
Market to Book	The shares outstanding multiplied by share price at the end of the quarter, all divided by the book value of equity	Compustat
Institutional%	Percentage of holdings from institutional investors	Thompson Reuters Institutional Holdings
News Sent	The average sentiment from news stories in the given time period	RavenPack
Q4	Dummy variable for announcements in the 4^{th} fiscal quarter	I/B/E/S

Appendix B LDA topic labels and word weight distributions

topic #	Topic	words and weights
0	recommendations	0.094*know + 0.079*make + 0.070*want + 0.057*got + 0.045*better + 0.044*play + 0.034*never + 0.026*level + 0.020*away + 0.018*shorting
1	past earnings info	0.255*er + 0.152*last + 0.146*time + 0.128*close + 0.060*open + 0.015*qtrs + 0.015*sp + 0.013*am + 0.013*pre + 0.011*fuel
2	speculation	0.124*sell + 0.048*many + 0.043*love + 0.038*give + 0.033*investor + 0.032*real + 0.027*set + 0.027*every + 0.023*value + 0.021*tell
3	attention to news	0.235*share + 0.128*|link + 0.031*interesting + 0.026*technoLny + 0.025*per + 0.024*picked + 0.024*director + 0.022*launch + 0.021*done + 0.021*corporation
4	noise glamour	0.044*gap + 0.039*bbry + 0.035*always + 0.033*business + 0.031*made + 0.029*minute + 0.028*pay + 0.026*-- + 0.026*more + 0.025*twitter
5	sec filing	0.239*new + 0.074*change + 0.049*sec + 0.049*form + 0.048*file + 0.038*reports + 0.036*filing + 0.035*events + 0.035*check + 0.033*passport
6	bullish news	0.083*bullish + 0.058*hour + 0.051*morning + 0.040*growth + 0.027*following + 0.026*article + 0.023*early + 0.018*name + 0.015*fda + 0.014*talking
7	tech charts	0.119*|link + 0.083*chart + 0.041*support + 0.034*q3 + 0.031*bounce + 0.031*show + 0.029*daily + 0.029*resistance + 0.026*bearish + 0.024*line
8	predict trends positive	0.160*go + 0.086*u + 0.056*up + 0.048*run + 0.045*higher + 0.039*sale + 0.035*:) + 0.032*ah + 0.026*ready + 0.024*lets
9	buying stock	0.050*point + 0.040*order + 0.024*finally + 0.024*trader + 0.022*baby + 0.022*remember + 0.021*mm + 0.020*talk + 0.020*off + 0.017*adding
10	earnings expectations	0.123*beat + 0.058*guidance + 0.054*miss + 0.041*squeeze + 0.036*the + 0.025*expectation + 0.015*will + 0.015*hot + 0.014*so + 0.014*q1

Appendix B (Cont.)

topic #	Topic	words and weights
11	upcoming earnings	0.328*earnings + 0.137*|link + 0.076*analyst + 0.074*release + 0.064*expect + 0.051*eps + 0.036*thursday + 0.025*wednesday + 0.017*tuesday + 0.017*anybody
12	attention	0.255*like + 0.169*look + 0.039*v + 0.031*trying + 0.020*find + 0.020*margin + 0.012*+003 + 0.012*broke + 0.012*agreement + 0.010*panic
13	noise	0.053*add + 0.045*: + 0.045*nq + 0.038*me + 0.022*ebola + 0.019*part + 0.018*car + 0.015*wtf + 0.014*andy + 0.013*ford
14	short sell	0.166*short + 0.049*lol + 0.045*right + 0.041*now + 0.040*way + 0.039*money + 0.029*getting + 0.027*bear + 0.024*around + 0.022*lower
15	bad news	0.090*news + 0.048*bad + 0.040*plug + 0.034*down + 0.033*action + 0.032*this + 0.029*thats + 0.026*nothing + 0.025*mean + 0.025*guess
16	questions about low	0.091*week + 0.085*next + 0.071*volume + 0.060*looking + 0.054*low + 0.041*anyone + 0.040*strong + 0.017*idea + 0.016*ago + 0.016*pretty
17	estimize	0.078*w + 0.071*symbol + 0.069*top + 0.061*estimize + 0.046*game + 0.039*tomorrow + 0.026*open + 0.019*covered + 0.015*ipad + 0.015*network
18	buy and hold	0.047*good + 0.039*see + 0.038*get + 0.037*going + 0.035*long + 0.031*market + 0.029*back + 0.028*one + 0.028*think + 0.024*would
19	rating	0.351*|link + 0.036*rating + 0.032*target + 0.030*blackberry + 0.030*pt + 0.019*corp + 0.016*capital + 0.013*bank + 0.012*security + 0.011*partner
20	cashtag	0.606*$cashtag + 0.256*#number + 0.018*call + 0.011*& + 0.009*put + 0.009*move + 0.007*option + 0.006*bought + 0.006*sold + 0.005*may
21	earnings news	0.113*|link + 0.095*quarter + 0.072*result + 0.051*announces + 0.043*financial + 0.039*first + 0.039*conference + 0.026*report + 0.024*bb + 0.022*must
22	insider trade	0.145*|link + 0.073*inc + 0.059*tonight + 0.056*million + 0.047*filed + 0.029*group + 0.025*update + 0.022*insider + 0.021*event + 0.020*ltd
23	technical	0.071*third + 0.035*upside + 0.028*surprise + 0.027*took + 0.024*wonder + 0.022*based + 0.021*later + 0.020*record + 0.019*offer + 0.018*load

Appendix B (Cont.)

topic #	Topic	words and weights
24	technical positive	0.296*- + 0.118*|link + 0.097*nice + 0.027*breakout + 0.025*second + 0.022*major + 0.020*due + 0.017*setup + 0.015*closing + 0.013*running
25	earnings speculation	0.064*= + 0.058*little + 0.049*ahead + 0.041*reason + 0.027*bit + 0.027*popular + 0.025*pain + 0.023*case + 0.021*q4 + 0.021*max
26	going long	0.069*holding + 0.061*it + 0.055*take + 0.049*well + 0.044*profit + 0.037*green + 0.034*huge + 0.031*bull + 0.027*gonna + 0.024*red
27	earnings expectations	0.193*price + 0.132*tomorrow + 0.079*announcement + 0.076*movement + 0.071*last + 0.069*qrtrs + 0.061*reporting + 0.037*friday + 0.033*|link + 0.013*po
28	market conditions	0.250*stock + 0.075*since + 0.063*drop + 0.050*trading + 0.049*continue + 0.049*er + 0.045*positive + 0.037*aftr + 0.030*past + 0.027*pm
29	advertising	0.059*great + 0.056*buying + 0.053*even + 0.042*selling + 0.036*monday + 0.027*in + 0.026*imo + 0.024*yet + 0.023*doesnt + 0.022*bottom
30	my trade	0.108*trade + 0.095*position + 0.059*number + 0.048*added + 0.025*investment + 0.025*recent + 0.024*retail + 0.022*date + 0.021*cap + 0.019*opening
31	analyst ratings2	0.124*eps + 0.097*estimate + 0.083*|link + 0.080*consensus + 0.080*rev + 0.073*report + 0.068*compared + 0.055*wall + 0.048*published + 0.047*estimize
32	discussion	0.412*@mention + 0.180*buy + 0.022*also + 0.020*bell + 0.016*thanks + 0.015*street + 0.012*too + 0.010*folk + 0.010*nasdaq + 0.009*told
33	stock pop	0.089*keep + 0.075*guy + 0.067*pop + 0.032*statement + 0.030*is + 0.030*anything + 0.027*ownership + 0.027*acquisition + 0.021*mkt + 0.020*late
34	apple	0.063*apple + 0.055*watch + 0.037*ceo + 0.036*|link + 0.024*work + 0.022*product + 0.020*system + 0.020*co + 0.017*billion + 0.016*store
35	earnings report	0.338*today + 0.085*revenue + 0.065*q314 + 0.041*potential + 0.030*block + 0.028*full + 0.020*est + 0.016*report: + 0.014*llc + 0.014*signal

Appendix B (Cont.)

<table>
<tr><th>topic #</th><th>Topic</th><th>words and weights</th></tr>
<tr><td>36</td><td>announcement</td><td>0.117*high + 0.062*quarterly + 0.044*cash + 0.043*dividend + 0.031*hard + 0.029*interest + 0.027*close + 0.026*range + 0.026*dec + 0.024*management</td></tr>
<tr><td>37</td><td>momentum</td><td>0.038*energy + 0.037*deal + 0.028*mobile + 0.025*head + 0.025*help + 0.024*plan + 0.018*cost + 0.018*key + 0.017*momentum + 0.017*global</td></tr>
<tr><td>38</td><td>trading results</td><td>0.176*day + 0.051*another + 0.048*break + 0.037*hit + 0.036*stop + 0.036*month + 0.034*end + 0.031*again + 0.024*loss + 0.023*gain</td></tr>
<tr><td>39</td><td>technical news</td><td>0.066*| + 0.057*tickerchirp + 0.049*tweet + 0.034*trend + 0.032*average + 0.027*risk + 0.026*via + 0.025*possible + 0.024*current + 0.023*read</td></tr>
</table>

9. Tables

Table 1 Descriptive Statistics

variable	N	mean	variance	p10	p50	p90	p99
$AR_{-10,-2}$	12,265	-0.004	0.005	-0.069	-0.004	0.060	0.201
$AR_{-1,1}$	12,265	0.000	0.016	-0.095	0.000	0.094	0.245
$AR_{2,60}$	12,265	-0.014	0.034	-0.215	-0.011	0.156	0.566
Ln(#Users)$_{-10,-2}$	12,265	1.946	1.298	0.693	1.609	3.401	5.908
Ln(#News Stories) $_{-10,2}$	12,265	5.270	1.456	4.025	5.130	6.739	8.968
Ln(#Retail EDGAR IP) $_{-10,2}$	12,265	2.965	7.283	0.000	4.277	6.011	7.547
SM Sent$_{-10,2}$	12,265	0.012	0.018	-0.144	0.000	0.189	0.394
News Sent$_{-10,2}$	12,265	0.016	0.000	-0.006	0.016	0.036	0.058
SUE	12,265	0.0002	0.0001	-0.005	0.0004	0.0056	0.0387
Market Cap	12,265	21.305	3.148	19.106	21.206	23.698	25.813
Ln(#Analysts)	12,265	2.151	0.498	1.099	2.197	3.091	3.497
Market to Book	12,265	4.036	42.848	0.986	2.574	8.642	46.364
Institutional %	12,265	0.622	0.052	0.306	0.661	0.850	1.007
Q4	12,265	0.2485	0.1868	0	0	1	1

This table provides the descriptive statistics for the main results in this study.

Table 2 Fama-French 48 Industry Classification

Industry Classification	Count	Industry Classification	Count
Agriculture	15	Machinery	327
Aircraft	71	Measuring and Control Equipment	190
Almost Nothing	51	Medical Equipment	320
Apparel	74	Non-Metallic and Industrial Metal Mining	53
Automobiles and Trucks	240	Personal Services	135
Banking	844	Petroleum and Natural Gas	558
Beer & Liquor	32	Pharmaceutical Products	761
Business Services	1,488	Precious Metals	31
Business Supplies	135	Printing and Publishing	74
Candy & Soda	39	Real Estate	56
Chemicals	283	Recreation	85
Coal	23	Restaurants, Hotels, Motels	231
Communication	261	Retail	289
Computers	294	Rubber and Plastic Products	75
Construction	162	Shipbuilding, Railroad Equipment	37
Construction Materials	139	Shipping Containers	28
Consumer Goods	135	Steel Works Etc.	139
Defense	16	Textiles	25
Electrical Equipment	124	Tobacco Products	36
Electronic Equipment	637	Trading	438
Entertainment	128	Transportation	310
Fabricated Products	21	Utilities	447
Food Products	134	Wholesale	311
Healthcare	205	*No SIC or not in Fama-French 48	1,297
Insurance	461	**Total**	**12,265**

This table presents the distribution of the sample for the main results by Fama-French 48 industry classification.

Table 3 Social media attention and overvaluation

$$AR_{i,j,t} = \alpha + \boldsymbol{\beta_1 Ln(\#Users)_{i,j,[-10,-2]}} + \sum_k \beta_k Control_{k,i,j} + e_{i,j,t}$$

	$AR_{-10,-2}$	$AR_{-1,1}$	$AR_{2,60}$
Constant	-0.125***	-0.109***	-0.207***
	(-10.01)	(-4.81)	(-6.94)
***Ln(#Users)**$_{-10,2}$*	**0.004*****	**-0.002****	**-0.010*****
	(3.24)	**(-2.02)**	**(-4.69)**
Ln(#News Stories) $_{-10,2}$	-0.002***	-0.002***	0.001
	(-3.40)	(-2.65)	(0.83)
Ln(#Retail EDGAR IP) $_{-10,2}$	-0.000	-0.001**	0.000
	(-0.86)	(-2.32)	(0.17)
SM Sent$_{-10,2}$	0.036***	0.026***	0.024
	(6.32)	(3.63)	(1.64)
News Sent$_{-10,2}$	0.334***	0.220***	0.076
	(8.10)	(4.65)	(0.78)
AR$_{-10,-2}$		-0.059***	-0.084**
		(-3.62)	(-2.39)
SUE		1.802***	0.130
		(10.77)	(0.47)
Market Cap	0.006***	0.007***	0.011***
	(8.73)	(5.52)	(6.40)
Ln(#Analysts)	-0.010***	-0.009***	-0.014***
	(-7.20)	(-5.03)	(-3.58)
Market to Book	-0.000	-0.000**	0.000
	(-0.93)	(-2.40)	(0.37)
Institutional %	0.009***	-0.002	0.004
	(2.86)	(-0.47)	(0.47)
Q4	0.013***	0.006***	-0.003
	(9.14)	(3.04)	(-0.78)
Observations	12,265	12,265	12,265
Adjusted R^2	0.034	0.034	0.008
Firm SE cluster	yes	yes	yes

This table presents of the results to test hypothesis H1. Models are run with clustered standard errors. *, **, *** indicate significance at p < 0.10, p < 0.05, and p < 0.01 levels, respectively (using two-tailed tests). The t-statistic is shown in parenthesis.

Table 4 Twitter vs. StockTwits (S&P 1500)

	$AR_{2,60}$	$AR_{2,60}$
Constant	-0.184***	-0.224***
	(-3.07)	(-3.45)
Ln(#StockTwits Users)$_{-10,-2}$	**-0.008***	
	(-1.87)	
Ln(#Twitter Users)$_{-10,-2}$		**-0.007*****
		(-2.59)
Ln(#News Stories)$_{-10,-2}$	0.001	0.001
	(0.15)	(0.29)
Ln(#Retail EDGAR IP)$_{-10,-2}$	0.000	0.000
	(0.36)	(0.27)
StockTwits Sent$_{-10,-2}$	0.024	
	(0.72)	
Twitter Sent$_{-10,-2}$		0.108***
		(3.31)
News Sent$_{-10,-2}$	0.070	0.062
	(0.35)	(0.31)
AR$_{-10,-2}$	-0.120	-0.117
	(-1.41)	(-1.39)
Market Cap	0.011***	0.013***
	(3.24)	(3.51)
Ln(#Analysts)	-0.026***	-0.026***
	(-3.31)	(-3.38)
Market to Book	0.000	0.000
	(0.61)	(0.47)
Institutional %	0.002	0.002
	(0.13)	(0.13)
Q4	-0.007	-0.006
	(-0.48)	(-0.42)
Observations	2,962	2,962
Adjusted R^2	0.005	0.010
Firm SE cluster	yes	yes

This table uses 45 million posts of Twitter data collected between August 2015 and December 2015. This set includes 99.9% of the posts from Twitter users that mentioned the cashtag of a company in the S&P 1500. Models are run with clustered standard errors. *, **, *** indicate significance at p < 0.10, p < 0.05, and p < 0.01 levels, respectively (using two-tailed tests). The t-statistic is shown in parenthesis.

Table 5 Social media attention and order imbalance

$$OIMB_{i,j,t} = \alpha + \boldsymbol{\beta_1 Ln(\#Users)_{i,j,[-10,-2]}} + \sum_k \beta_k Control_{k,i,j} + e_{i,j,t}$$

	Retail $OIMB_{-10,-2}$	Overall $OIMB_{-10,-2}$	Retail $OIMB_{-1,1}$	Overall $OIMB_{-1,1}$	Retail $OIMB_{2,60}$	Overall $OIMB_{2,60}$
Constant	0.016	-0.094***	0.057	-0.023	-0.073*	-0.097***
	(0.40)	(-8.38)	(1.25)	(-1.28)	(-1.91)	(-9.29)
Ln(#Users) $_{-10,2}$	**0.021*** **	**0.002*** **	**0.017*** **	**0.001**	**0.020*** **	**-0.000**
	(9.60)	**(2.94)**	**(7.02)**	**(0.98)**	**(9.66)**	**(-0.05)**
Ln(#News Stories) $_{-10,2}$	0.008**	-0.002**	0.008**	0.001	0.005*	-0.000
	(2.44)	(-2.33)	(2.36)	(0.95)	(1.77)	(-0.67)
Ln(#Retail EDGAR IP) $_{-10,2}$	0.002*	-0.000	0.002	-0.000	0.002*	-0.000*
	(1.69)	(-0.66)	(1.54)	(-0.71)	(1.66)	(-1.70)
SM Sent $_{-10,2}$	0.053***	0.014***	0.042**	0.003	0.048***	0.009*
	(2.97)	(2.67)	(2.06)	(0.36)	(3.00)	(1.90)
News Sent $_{-10,2}$	-0.012	0.018	-0.038	-0.225***	0.004	0.002
	(-0.10)	(0.49)	(-0.25)	(-3.65)	(0.04)	(0.07)
SUE			0.250	-0.202	-0.025	0.077
			(1.13)	(-1.49)	(-0.19)	(1.12)
Market Cap	-0.010***	0.004***	-0.011***	-0.000	-0.005**	0.003***
	(-4.19)	(6.01)	(-4.03)	(-0.25)	(-2.06)	(5.84)
Ln(#Analysts)	0.005	0.000	0.007	0.004**	0.001	0.003***
	(0.94)	(0.35)	(1.05)	(2.09)	(0.27)	(2.65)
Market to Book	0.001	0.000	0.000	0.000***	0.001	0.000**
	(1.57)	(0.32)	(0.48)	(2.86)	(1.32)	(2.02)
Institutional %	-0.084***	0.018***	-0.085***	0.029***	-0.111***	0.021***
	(-5.97)	(5.26)	(-5.68)	(5.14)	(-7.81)	(6.50)
Q4	0.005	0.000	0.008	-0.001	0.006	0.002
	(1.04)	(0.23)	(1.38)	(-0.51)	(1.53)	(1.44)
Observations	10,821	10,821	10,821	10,821	10,821	10,821
Adjusted R^2	0.031	0.017	0.019	0.009	0.043	0.037
Firm SE cluster	yes	yes	yes	yes	yes	yes

This table shows the relationship between order imbalance in the -10 to -2 days, -1 to 1 days, and 2 to 60 days relative to the earnings announcement for a given firm and attention on social media. Retail order imbalance is calculated using the algorithm in Boehmer et al. (2017). Overall order imbalance is calculated using the algorithm in Chakrabarty et al. (2007). Models are run with clustered standard errors. *, **, *** indicate significance at $p < 0.10$, $p < 0.05$, and $p < 0.01$ levels, respectively (using two-tailed tests). The t-statistic is shown in parenthesis.

Table 6 User Attention and Short-selling

$$AR_{i,j,t} = \alpha + \boldsymbol{\beta_1 Ln}(\#\boldsymbol{Users})_{\boldsymbol{i,j},[-\mathbf{10},-\mathbf{2}]} + \boldsymbol{\beta_2 Ln}(\#\boldsymbol{Topics})_{\boldsymbol{i,j},[-\mathbf{10},-\mathbf{2}]} + \sum_k \beta_k Control_{k,i,j} + e_{i,j,t}$$

	$\text{Short}_{-10,-2}$	$\text{Short}_{-1,1}$	$\text{Short}_{2,60}$
Constant	0.191***	0.098***	1.523***
	(12.60)	(11.15)	(9.91)
Log(#Users)$_{-10,2}$	**0.019*** **	**0.007*** **	**0.092*** **
	(3.45)	**(4.68)**	**(5.82)**
Log(#News Stories$_{-10,-2}$)	-0.000	0.000	0.006
	(-0.10)	(0.60)	(1.22)
Log(#Retail EDGAR IP)$_{-10,-2}$	-0.000	0.000	0.001
	(-0.02)	(0.72)	(0.30)
SM Sent$_{-10,-2}$	-0.005	-0.003	0.031
	(-0.64)	(-1.00)	(0.91)
News Sent$_{-10,-2}$	-0.047	-0.051*	-0.477
	(-0.57)	(-1.90)	(-1.42)
AR$_{-10,-2}$		0.017	0.114
		(1.34)	(1.00)
SUE		0.066	-0.170
		(1.09)	(-0.11)
Market Cap	-0.010***	-0.006***	-0.081***
	(-11.15)	(-10.25)	(-9.02)
Log(#Analysts)	0.009**	0.008***	0.099***
	(2.11)	(5.96)	(5.39)
Market to Book	-0.001**	-0.000***	-0.003***
	(-2.27)	(-2.77)	(-3.15)
Institutional %	-0.010**	-0.003	-0.106***
	(-2.01)	(-1.27)	(-3.09)
q4	0.002	0.000	0.004
	(0.69)	(0.55)	(0.60)
Observations	12,008	12,008	12,008
Adjusted R^2	0.109	0.148	0.130
Firm SE Cluster	yes	yes	yes

The dependent variables in this table are market-adjusted abnormal returns in the -10 to -2 days, -1 to 1 days, and 2 to 60 days relative to the earnings announcement. *Ln(#Topics)* is the entropy of the topics discussed on StockTwits in the -10 to -2 days relative to earnings. Models are run with clustered standard errors. *, **, *** indicate significance at $p < 0.10$, $p < 0.05$, and $p < 0.01$ levels, respectively (using two-tailed tests). The t-statistic is shown in parenthesis.

Table 7 Cross-sectional analysis

Panel A: Institutional Ownership

$$AR_{i,j,[2,60]} = \alpha + \boldsymbol{\beta_1 Ln(\#Users)_{i,j,[-10,-2]}} + \sum_k \beta_k Control_{k,i,j} + e_{i,j,t}$$

	Q1	Q2	Q3	Q4	Q5
Constant	-0.193***	-0.138*	-0.112	-0.138	-0.034
	(-2.98)	(-1.73)	(-1.11)	(-1.01)	(-0.35)
***Log(#Users)**-10,2*	**-0.014*** **	**-0.011** **	**-0.013*** **	**-0.003**	**-0.001**
	(-3.28)	**(-2.27)**	**(-2.91)**	**(-0.72)**	**(-0.26)**
Observations	2,453	2,453	2,453	2,453	2,453
Adjusted R^2	0.011	0.016	0.007	0.004	0.001
Controls	yes	yes	yes	yes	yes
Firm SE cluster	yes	yes	yes	yes	yes

Panel B: Analyst dispersion

	Q1	Q2	Q3	Q4	Q5
Constant	-0.007	-0.151**	-0.185***	-0.284***	-0.358***
	(-0.11)	(-2.28)	(-2.79)	(-4.06)	(-4.83)
***Log(#Users)**-10,2*	**-0.004**	**-0.005**	**-0.005**	**-0.010** **	**-0.018*** **
	(-0.96)	**(-1.09)**	**(-1.10)**	**(-2.25)**	**(-3.26)**
Observations	2,453	2,453	2,453	2,453	2,453
Adjusted R^2	0.004	0.002	0.006	0.012	0.019
Controls	yes	yes	yes	yes	yes
Firm SE cluster	yes	yes	yes	yes	yes

Panel C: Analyst following

	Q1	Q2	Q3	Q4	Q5
Constant	-0.090	-0.008	-0.178**	-0.281***	-0.366***
	(-0.99)	(-0.08)	(-2.14)	(-3.42)	(-4.34)
***Log(#Users)**-10,2*	**-0.010***	**-0.011** **	**-0.010** **	**-0.006**	**-0.013*** **
	(-1.76)	**(-2.01)**	**(-2.35)**	**(-1.58)**	**(-3.76)**
Observations	2,453	2,453	2,453	2,453	2,453
Adjusted R^2	0.003	0.009	0.003	0.010	0.054
Controls	yes	yes	yes	yes	yes
Firm SE cluster	yes	yes	yes	yes	yes

Table 7 (Cont.)

Panel D: Social media disagreement

	Q1	Q2	Q3	Q4	Q5
Constant	-0.049	-0.274***	-0.180***	-0.205***	-0.280***
	(-0.71)	(-4.12)	(-2.90)	(-2.80)	(-4.36)
Log(#Users)$_{-10,2}$	**0.003**	**-0.007***	**-0.013*****	**-0.009***	**-0.018*****
	(0.47)	**(-1.71)**	**(-3.31)**	**(-1.94)**	**(-3.03)**
Observations	2,453	2,453	2,453	2,453	2,453
Adjusted R^2	0.005	0.011	0.008	0.004	0.015
Controls	yes	yes	yes	yes	yes
Firm SE cluster	yes	yes	yes	yes	yes

Panel E: Social media and media sentiment

	SM+,M-	SM-,M+	SM+,M+	SM-,M-
Constant	-0.285	-0.212***	-0.190***	-0.168
	(-1.58)	(-5.14)	(-4.20)	(-1.59)
Log(#Users)$_{-10,2}$	**-0.012**	**-0.005**	**-0.013*****	**-0.004**
	(-1.09)	**(-1.56)**	**(-4.96)**	**(-0.49)**
Observations	817	5,071	5,317	1,038
Adjusted R^2	0.019	0.009	0.011	0.009
Controls	yes	yes	yes	yes
Firm SE cluster	yes	yes	yes	yes

This table presents the cross-sectional regression results that focus on limits to arbitrage, attention from analysts and size (all quintiles are sorted from low to high). Panel A shows the results based on quintile of institutional ownership. Panel B shows the results based on quintile of analyst dispersion. Panel C shows the results based on quintile of analyst following. Panel D shows the results based on quintile of social media disagreement. Panel E is divided by the sentiment on social media and the sentiment in the news. SM+ indicates positive sentiment on social media, M+ indicates positive sentiment in media articles. The sentiment divisions are made on sentiment in the -10 to -2 days relative to each quarterly earnings announcement in the sample. Models are run with clustered standard errors. *, **, *** indicate significance at $p < 0.10$, $p < 0.05$, and $p < 0.01$ levels, respectively (using two-tailed tests). The t-statistic is shown in parenthesis.

Table 8 Buying bulls

$$AR_{i,j,t} = \alpha + \boldsymbol{\beta_1 Ln(\#Users)_{i,j,[-10,-2]}} + \sum_k \beta_k Control_{k,i,j} + e_{i,j,t}$$

	$AR_{-10,-2}$		$AR_{2,60}$	
	Pos SM Sent	Neg SM Sent	Pos SM Sent	Neg SM Sent
Constant	-0.117***	-0.139***	-0.214***	-0.210***
	(-6.30)	(-7.86)	(-4.75)	(-5.22)
Ln(#Users)-10,-2	**0.006*** **	**-0.000**	**-0.013*** **	**-0.005**
	(4.21)	**(-0.03)**	**(-4.59)**	**(-1.57)**
Ln (#News Stories) -10,2	-0.003***	-0.001	0.002	0.001
	(-3.29)	(-1.62)	(0.74)	(0.37)
Ln(#Retail EDGAR IP) -10,2	-0.000	0.000	0.001	-0.001
	(-1.09)	(0.10)	(0.79)	(-0.82)
SM Sent-10,2	0.027**	0.032**	0.060**	0.003
	(2.43)	(2.28)	(1.97)	(0.10)
News Sent-10,2	0.306***	0.340***	0.021	0.157
	(5.08)	(6.29)	(0.14)	(1.34)
AR-10,-2			-0.068	-0.097*
			(-1.35)	(-1.95)
SUE			0.109	0.108
			(0.28)	(0.29)
Market Cap	0.006***	0.007***	0.011***	0.010***
	(5.54)	(7.13)	(4.32)	(4.72)
Log(#Analysts)	-0.008***	-0.011***	-0.013**	-0.015***
	(-4.16)	(-5.44)	(-2.21)	(-3.09)
Market to Book	-0.000	0.000	-0.000	0.001*
	(-1.61)	(0.49)	(-0.52)	(1.82)
Institutional %	0.010**	0.010**	-0.004	0.014
	(2.12)	(2.18)	(-0.37)	(1.15)
Q4	0.016***	0.010***	-0.018***	0.013***

Table 8 (Cont.)

	(7.00)	(5.98)	(-3.19)	(2.82)
Observations	6,144	6,121	6,144	6,121
Adjusted R^2	0.031	0.036	0.009	0.008
Firm SE cluster	yes	yes	yes	yes

The dependent variables in this table are market-adjusted abnormal returns in the -10 to -2 days, -1 to 1 days, and 2 to 60 days relative to the earnings announcement. The sample is divided by average sentiment on social media. Models are run with clustered standard errors. *, **, *** indicate significance at $p < 0.10$, $p < 0.05$, and $p < 0.01$ levels, respectively (using two-tailed tests). The t-statistic is shown in parenthesis.

Table 9 Within-network dissemination and herding

$$AR_{i,j,t} = \alpha + \boldsymbol{\beta_1 Ln(\#Users)_{i,j,[-10,-2]}} + \boldsymbol{\beta_2 Ln(\#Top\ Followed\ Users)_{i,j,[-10,-2]}} + \sum_k \beta_k Control_{k,i,j} + e_{i,j,t}$$

	$AR_{-10,-2}$	$AR_{-1,1}$	$AR_{2,60}$
Constant	-0.096***	-0.092**	-0.191***
	(-5.67)	(-2.02)	(-4.75)
Ln(#Top Followed Users)$_{-10,2}$	**0.008*** **	**0.004**	**0.003**
	(4.88)	**(1.09)**	**(0.68)**
Ln(#Users)$_{-10,2}$	**0.001**	**-0.005** **	**-0.007** **
	(0.56)	**(-2.27)**	**(-2.22)**
Ln(#News Stories)$_{-10,2}$	-0.003***	-0.002	0.001
	(-3.38)	(-1.53)	(0.28)
Ln(#Retail EDGAR IP)$_{-10,2}$	-0.000	-0.001	0.000
	(-1.22)	(-1.40)	(0.28)
SM Sent$_{-10,2}$	0.037***	0.030**	0.007
	(4.31)	(2.51)	(0.35)
News Sent$_{-10,2}$	0.376***	0.225***	0.094
	(6.35)	(3.43)	(0.74)
AR$_{-10,-2}$		-0.057**	-0.133***
		(-2.45)	(-2.90)
SUE		1.985***	0.063
		(5.72)	(0.15)
Market Cap	0.005***	0.006***	0.010***
	(5.07)	(2.68)	(4.63)
Ln(#Analysts)	-0.009***	-0.010***	-0.019***
	(-4.74)	(-4.00)	(-3.52)
Market to Book	-0.000	-0.000**	-0.000
	(-1.23)	(-2.42)	(-0.06)
Institutional %	0.011**	-0.002	0.014
	(2.45)	(-0.22)	(1.05)
Q4	0.012***	0.011***	-0.009*
	(6.32)	(3.33)	(-1.87)
Observations	6,514	6,514	6,514
Adjusted R^2	0.041	0.022	0.008
Firm SE cluster	yes	yes	yes

The dependent variables in this table are market-adjusted abnormal returns in the -10 to -2 days, -1 to 1 days, and 2 to 60 days relative to the earnings announcement. *Ln(#Top Followed Users)$_{-10,-2}$* is the natural log of the number of unique active users in the set of the 1000 top-followed users by year in the -10 to -2 days relative to earnings. Models are run with clustered standard errors. *, **, *** indicate significance at $p < 0.10$, $p < 0.05$, and $p < 0.01$ levels, respectively (using two-tailed tests). The t-statistic is shown in parenthesis.

Table 10 User-generated information and overconfidence

$$AR_{i,j,t} = \alpha + \boldsymbol{\beta_1 Ln(\#Users)_{i,j,[-10,-2]}} + \boldsymbol{\beta_2 Ln(\#Topics)_{i,j,[-10,-2]}} + \sum_k \beta_k Control_{k,i,j} + e_{i,j,t}$$

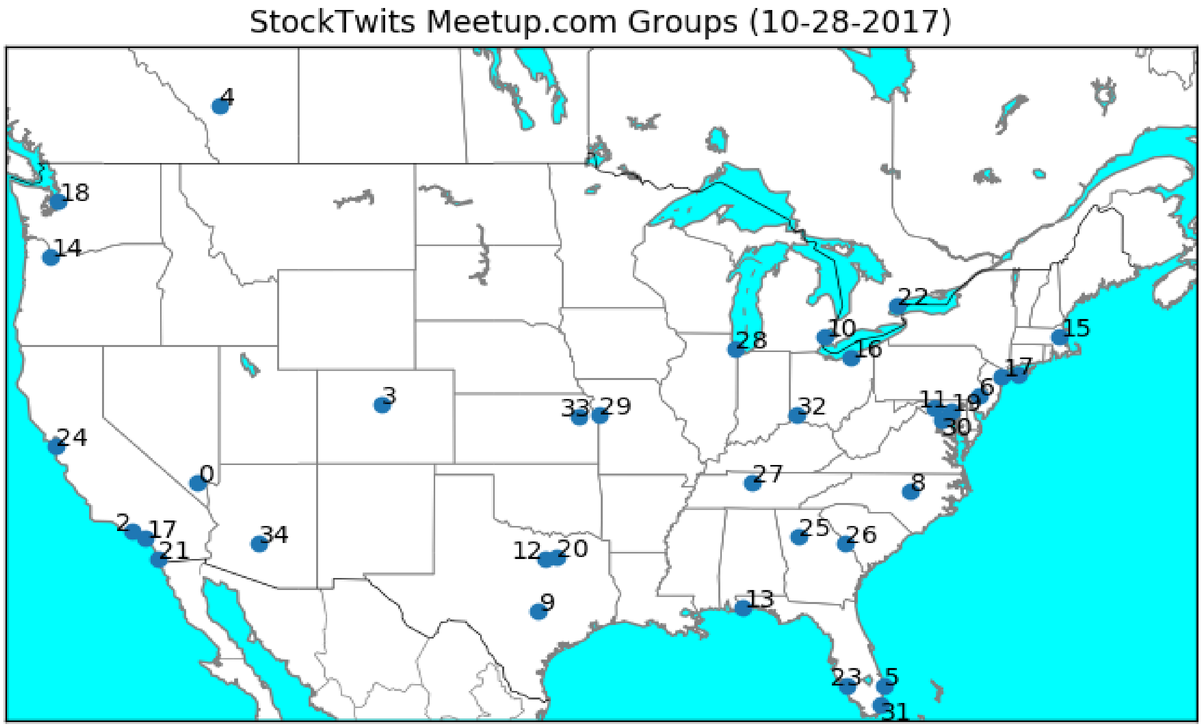
	$AR_{-10,-2}$	$AR_{-1,1}$	$AR_{2,60}$
Constant	-0.125***	-0.109***	-0.207***
	(-10.00)	(-4.81)	(-6.94)
Ln(#Topics)	**-0.001**	**-0.004***	**-0.013****
	(-0.54)	**(-1.79)**	**(-3.32)**
Ln(#Users) $_{-10,2}$	**0.004*****	**0.001**	**-0.001**
	(3.10)	**(0.44)**	**(-0.39)**
Ln(#News Stories) $_{-10,2}$	-0.002***	-0.002***	0.002
	(-3.39)	(-2.63)	(0.93)
Ln(#Retail EDGAR IP) $_{-10,2}$	-0.000	-0.001**	0.000
	(-0.87)	(-2.32)	(0.17)
SM Sent $_{-10,2}$	0.036***	0.028***	0.029*
	(6.33)	(3.89)	(1.93)
News Sent $_{-10,2}$	0.333***	0.216***	0.061
	(8.08)	(4.56)	(0.62)
AR $_{-10,-2}$		-0.059***	-0.084**
		(-3.63)	(-2.38)
SUE		1.801***	0.110
		(10.76)	(0.41)
Market Cap	0.006***	0.007***	0.010***
	(8.77)	(5.56)	(6.20)
Ln(#Analysts)	-0.010***	-0.009***	-0.013***
	(-7.17)	(-4.96)	(-3.42)
Market to Book	-0.000	-0.000**	0.000
	(-0.91)	(-2.29)	(0.51)
Institutional %	0.009***	-0.003	0.003
	(2.83)	(-0.56)	(0.32)
Q4	0.013***	0.007***	-0.002
	(9.10)	(3.10)	(-0.41)
Observations	12,265	12,265	12,265
Adjusted R^2	0.034	0.034	0.008
Firm SE cluster	yes	yes	yes

The dependent variables in this table are market-adjusted abnormal returns in the -10 to -2 days, -1 to 1 days, and 2 to 60 days relative to the earnings announcement. *Ln(#Topics)* is the entropy of the topics discussed on StockTwits in the -10 to -2 days relative to earnings. Models are run with clustered standard errors. *, **, *** indicate significance at $p < 0.10$, $p < 0.05$, and $p < 0.01$ levels, respectively (using two-tailed tests). The t-statistic is shown in parenthesis.

Table 11 Grainger-causality tests

Full Timespan (2014, 2015)			
Variable	Grainger Causes	Chi-Squared	Probability>Chi-Squared
ΔLn(#Users)	*Return*	36.461***	0.000
Return	*ΔLn(#Users)*	11.928***	0.001
2 to 60 Days After Earnings Announcement			
Variable	Grainger Causes	Chi-Squared	Probability>Chi-Squared
ΔLn(#Users)	*Return*	0.114	0.736
Return	*ΔLn(#Users)*	5.483**	0.019

10. Figures



index	place	#members	index	place	#members
0	Las Vegas	282	18	Seattle	768
1	New York City	1,690	19	Baltimore	371
2	Santa Monica	650	20	Dallas	623
3	Colorado	449	21	San Diego	311
4	Calgary	370	22	Toronto	734
5	West Palm Beach	339	23	Fort Meyers	164
6	Philadelphia	394	24	San Francisco	378
7	Long Island	274	25	Atlanta	330
8	Raleigh-Durham	223	26	Augusta	348
9	Austin	352	27	Nashville	129
10	Detroit	497	28	Chicago	477
11	Frederick	35	29	Kansas City	507
12	Fort Worth	272	30	Washington DC	215
13	Pensacola	301	31	Miami	177
14	Portland	21	32	Cincinnati	138
15	Boston	693	33	Topeka	61
16	Cleveland	501	34	Phoenix	92
17	Irvine	271	Total	World	13,437

Figure 1. StockTwits meetups.

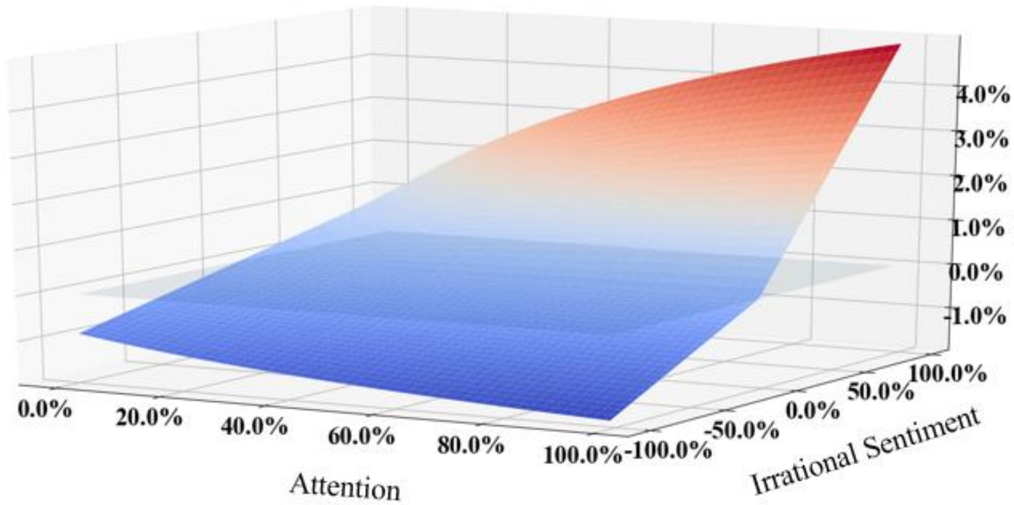


Figure 2.1. Simulation of mispricing using equilibrium point in Kelly and Tetlock (2013). I use the equilibrium price for short-short selling constrained investors without disagreement:

$$P_{constrained} = \frac{F+s_I+ks_t}{2+k} + \frac{(\gamma_B+[\theta+I(m>0)(1-\theta)]\alpha\gamma_C)m}{(\gamma_A+\gamma_B+\alpha\gamma_C)/(2+k)}.$$

I use $\gamma_A = 10, \gamma_B = 1, \gamma_C = 10, \theta = 0.3$, and vary sentiment, m and attention, α. The variable γ is the risk tolerance of each investor group in {A, B, C}. The variable α is the level of attention from investors in group A. The variable θ represents the percent of the population in the set A that is not short-selling constrained. I normalize the fundamental price and scale by a factor of 10.

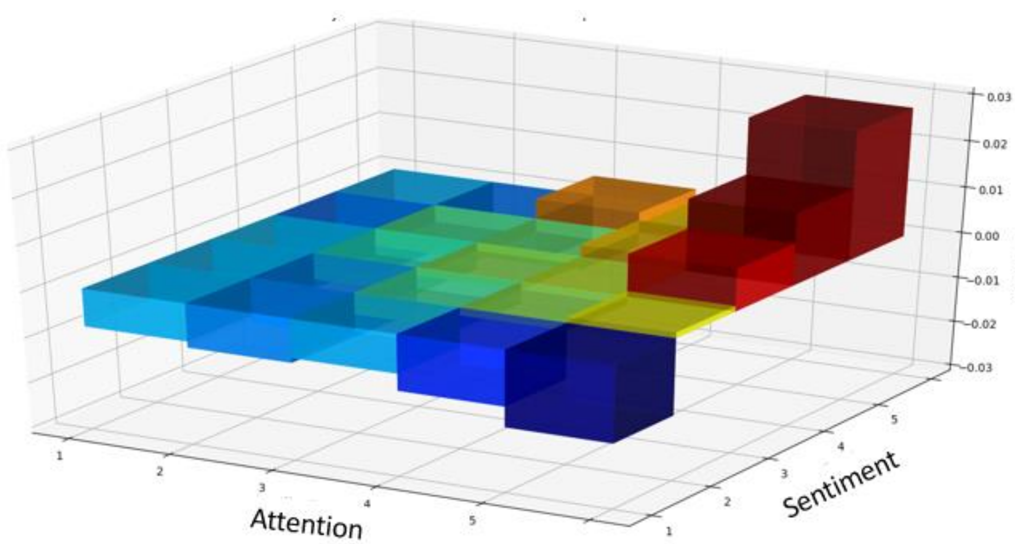


Figure 2.2. Actual sentiment and attention by quintile of sentiment and attention on StockTwits.

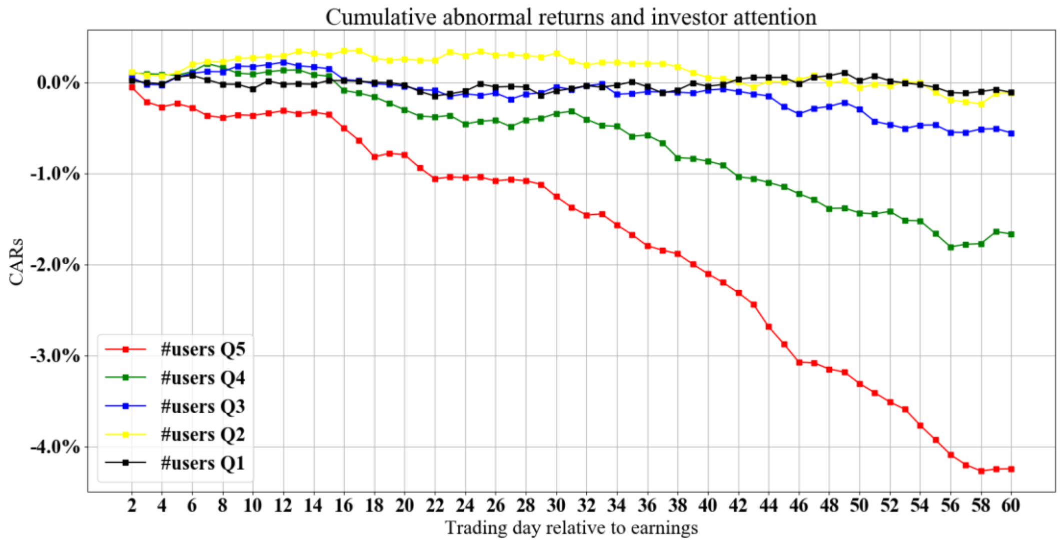


Figure 3. Post-earnings cumulative abnormal return drift by quintile of StockTwits attention

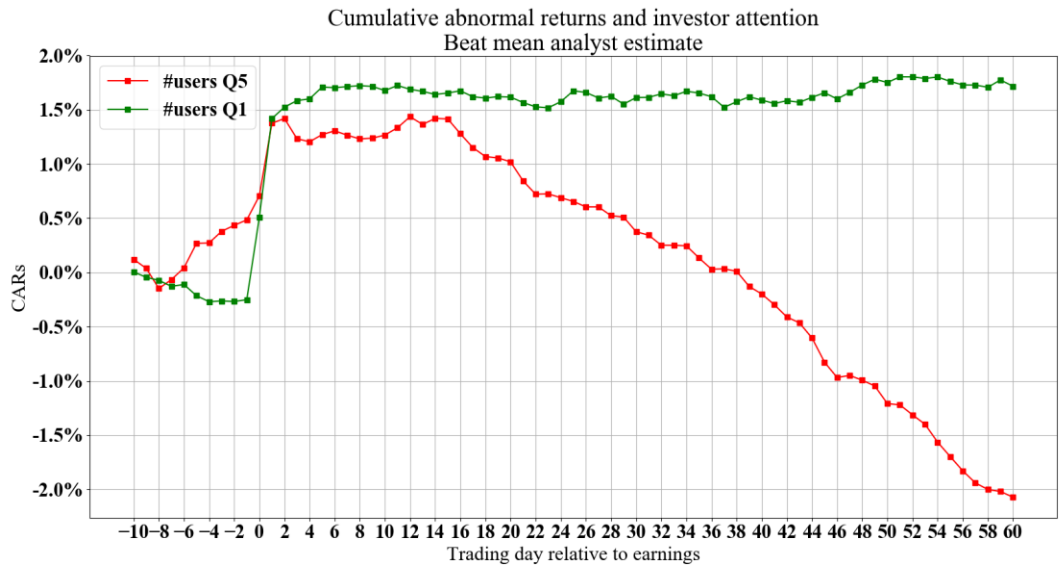


Figure 4.1. Extreme quintile cumulative abnormal returns for companies that had net positive sentiment and beat analyst expectations.

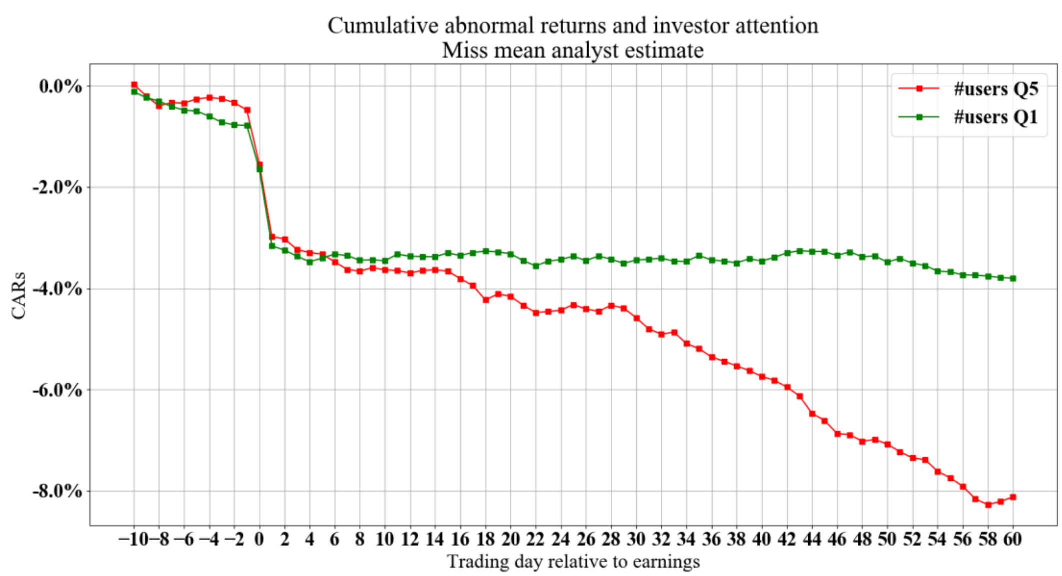


Figure 4.2. Extreme quintile cumulative abnormal returns for companies that had net positive sentiment and missed analyst expectations.

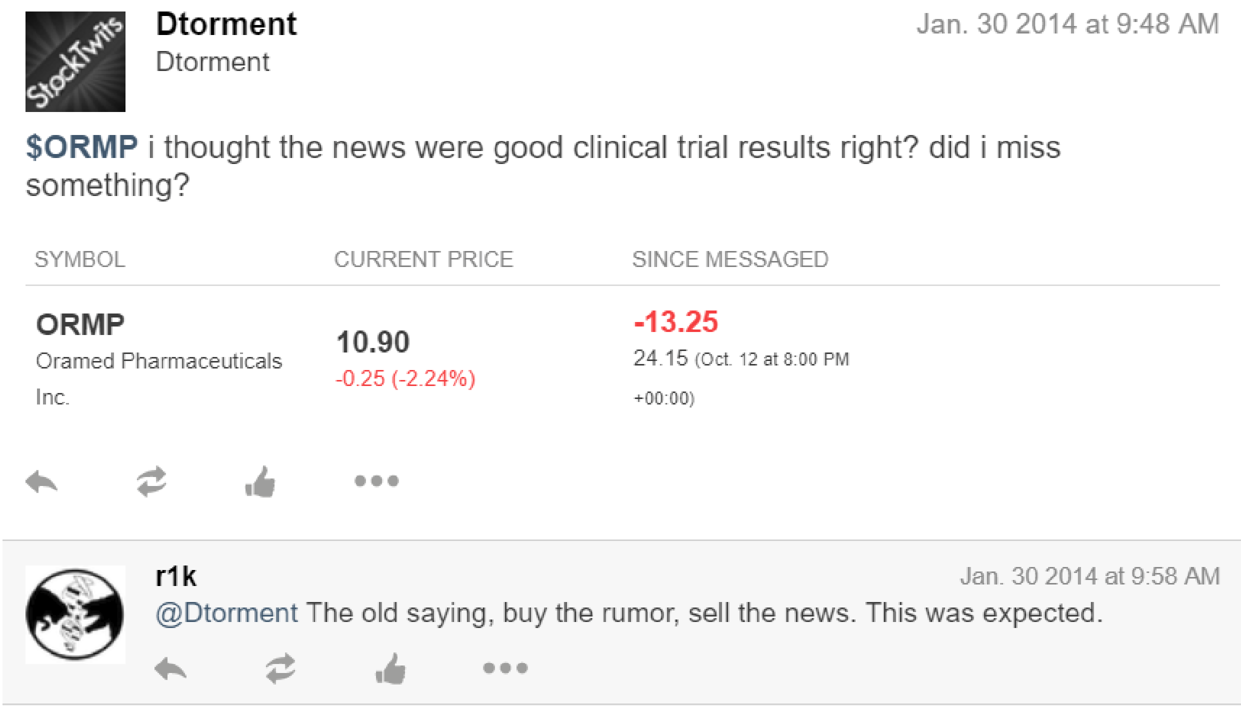


Figure 5: StockTwits post example.