The Impact of Public Disclosure on Information Asymmetry between Sophisticated and Unsophisticated Investors: Evidence from an Investor Social Media Network

Donal Byard

Baruch College- City University of New York

Yakun Wang

New York University – Shanghai
Southwestern University of Finance and Economics

October 2016

Abstract

We collect a unique dataset from StockTwits.com, the most popular social media networks dedicated to the discussion of stock investments, to examine the impact of public disclosure on information asymmetry between sophisticated and unsophisticated investors. We apply Naïve-Bayes textual analysis method to extract information from about 20 million tweets posted by investors of various sophistications and develop a daily measure of the degree of information asymmetry between sophisticated and unsophisticated investors. Using sporadic management forecasts as a research setting, we find that: 1) information asymmetry between these two classes of investors increases in the short term (about a week) after public disclosures-specifically, sporadic management forecasts; 2) information asymmetry decreases in the long run (roughly a week) after public disclosures; and 3) more precise public disclosures result in a smaller short-term increase (and a larger long-term decrease) in information asymmetry.

Keywords: Social Media; Information Asymmetry; Investor Sophistication

We could like to sincerely thank Orie Barron, Edward Li, Hagit Levy, Carol Marquardt, Lin Peng, Joseph Weintrop, Yong Yu, and seminar participants at Baruch College – CUNY, New York University, Renmin University, Wuhan University for their helpful comments. We would also like to gratefully acknowledge the support from Stocktwits.com, especially Chris Corriveau for providing access to the database and API service.

The Impact of Public Disclosure on Information Asymmetry between Sophisticated and Unsophisticated Investors: Evidence from an Investor Social Media Network

Abstract

We collect a unique dataset from StockTwits.com, the most popular social media networks dedicated to the discussion of stock investments, to examine the impact of public disclosure on information asymmetry between sophisticated and unsophisticated investors. We apply Naïve-Bayes textual analysis method to extract information from about 20 million tweets posted by investors of various sophistications and develop a daily measure of the degree of information asymmetry between sophisticated and unsophisticated investors. Using sporadic management forecasts as a research setting, we find that: 1) information asymmetry between these two classes of investors increases in the short term (about a week) after public disclosures-specifically, sporadic management forecasts; 2) information asymmetry decreases in the long run (roughly a week) after public disclosures; and 3) more precise public disclosures result in a smaller short-term increase (and a larger long-term decrease) in information asymmetry.

Keywords: Social Media; Information Asymmetry; Investor Sophistication

1. Introduction

The stock market consists of investors of varying levels of sophistication. Differences in investors' ability in collecting and analyzing information can result in an uneven informational playing field (i.e. information asymmetry) between sophisticated and unsophisticated investors. The resulting adverse selection problem can have significant economic consequences regarding the increase in firms' cost of capital (e.g. Watts and Zimmerman 1979, Grossman and Stiglitz 1980, Diamond and Verrecchia 1991, Leuz and Verrecchia 2000). Additionally, sophisticated investors' information advantage can enable them to profit from trading against unsophisticated investors, thus compromising the fairness of the stock market (see Lev 1988 for a discussion). As a result of such concerns, securities regulators have adopted many policies to "level the informational playing field" across all investors, including the adoption of disclosure policies designed to ensure that all investors have equal access to public disclosures.[1]

However, despite the large body of theoretical literature investigating the impact of public disclosure on information asymmetry, to date, there is little direct evidence as to whether or not public disclosures serve to "level the informational playing field" between sophisticated and unsophisticated investors. In part, this is because we lack direct measures of the degree of information asymmetry between these two classes of investors. This paper

[1] Based upon the Securities and Exchange Act of 1933, the Securities and Exchange Commission (SEC) has enacted a disclosures regime that is designed to level the informational playing field among investors. Indeed, in 1998, former Chairman of the SEC Arthur Levitt remarked that "the statutes establishing our regulatory system championed the idea of the level playing field." Similarly, regarding the purpose of the U.S. mandatory disclosure regime, Lev (1988) observed that the "adverse consequences of inequity can be mitigated by a public policy mandating the disclosure of financial information in order to reduce information asymmetries." More recently, the SEC adopted regulation Fair Disclosure (FD) aimed at "leveling the informational playing field" among different types of investors by curbing firms' practice of selective disclosing to a subset of investors.

seeks to address this gap in the literature. In this paper, we develop a direct new measure of information asymmetry between sophisticated and unsophisticated investors based on the content they posted on social media; and using this new measure we test the effect of unanticipated firm-provided public disclosures on the degree of information asymmetry between sophisticated and unsophisticated investors.

The role of public disclosures in leveling the informational playing field between sophisticated and unsophisticated investors is complex. On the one hand, public disclosures can reveal to unsophisticated investors (i.e. uninformed traders) information that was previously possessed only by sophisticated investors (i.e. informed traders), thus leveling the informational playing field between sophisticated and unsophisticated investors (Kim and Verrecchia 1991, Atiase and Bamber 1994). On the other hand, studies that focus on sophisticated investors' greater ability to analyze public disclosures (e.g., Kim and Verrecchia 1994, 1997, Bamber et al. 1999) suggest that, compared to unsophisticated investors, sophisticated investors can better interpret the information provided by public disclosures to generate new private information. In this case, public disclosures can potentially serve to *increase* the information asymmetry between sophisticated and unsophisticated investors, at least in the short run (Kim and Verrecchia 1997). To date, however, we have little direct empirical evidence that speaks to whether or not public disclosures serve to level the informational playing field between sophisticated and unsophisticated investors, in large part because we lack a direct measure of the degree of information asymmetry between sophisticated and unsophisticated investors.[2]

[2] Trade size has been widely used as a proxy for investor sophistications. However, recent studies (Cready et al. 2014 and Frazzini et al. 2012) call into question the validity of using trade size to infer investor sophistications, by showing that highly professional investors actively and strategically separate large trading orders into small sized trades, making trade-size a noisy (even biased) proxy of investor sophistication.

In this study, we use social media data to develop a new and direct measure of information asymmetry (see below) between sophisticated and unsophisticated investors. Using this new measure, we examine the impact of public disclosures on the degree of information asymmetry between these two classes of investors. We focus on sporadic management forecasts (management forecasts that are not made within +/- 3-days of a scheduled earnings announcements, or rendered on a periodic basis) for the following reasons: First, these forecasts are usually unanticipated by investors, therefore eliminating the impact of private information search activities specifically in anticipation of an scheduled disclosure (see Kim and Verrecchia 1991b) on information asymmetry, providing for a cleaner setting to test the impact of public disclosure per se on the information asymmetry between sophisticated and unsophisticated investors. Second, firms issue management forecasts of varying levels of precision (i.e., point vs. range forecasts, see Baginski et al. 1993), allowing for empirical tests of how public disclosures of varying precision affect information asymmetry between sophisticated and unsophisticated investors.[3]

This new measure is constructed using investor-generated statements posted on StockTwits.com, the most popular social media network dedicated to investment activity in the U.S.[4] On this platform investors can post short discussions (of up to 140 characters), known as "tweets," to express their belief in the investment potential of particular stocks. Because these tweets are posted directly by individual investors, researchers can potentially

[3] In Contrast, it's more difficult to identify the level of precision for earnings announcements.

[4] As of July of 2014, StockTwits has more than 600,000 active users, and an average monthly volume of more than 3 million tweets. And according to StockTwits, "the posts are viewed by an audience of over 40 million across the financial web and social media platforms."- From introduction of StockTwits.com (http://stocktwits.com/about).

use measures of the content of these tweets to proxy for individual investors' beliefs about the investment value of specific stocks. Additionally, since tweets post almost continuously, researchers can use this data to develop relatively high-frequency (i.e., daily) measures of investors' beliefs regarding a particular stock. Applying textual analysis to a sample of 20 million tweets posted to StockTwits.com between 2009 and 2013 by both sophisticated and unsophisticated investors, we extract the information content (positive, negative, or neutral) of these tweets and develop measures of the average beliefs of both sophisticated and unsophisticated investors. Specifically, based upon users' sophistication levels, on a daily basis we calculate the mean values of the information sets of both sophisticated and unsophisticated investors; then, using this data, we construct a measure of the information asymmetry between sophisticated and unsophisticated investors as the difference between their average information sets.

Before applying the new measure, we conduct a series of tests to establish its construct validity. First, we examine the accuracy of Naïve-Bayes text classification using a 10-fold cross validation method. Thanks to the relatively large number of training sample we have acquired from StockTwits (about 1.1 million tweets) that are labeled "positive" or "negative" by the authors of the tweets, we are able to achieve an average classification accuracy of 88%. [5] Second, we confirm the validity of aggregated firm-day level measures

[5] For our main test we use Naïve-Bayes (NB) approach to classify the information of posts, our training material includes about 1.1 millions tweets whose tone are labeled by authors of the posts. StockTwits facilitate a system that when posting a tweet, user can voluntarily choose to disclosure her/his opinion on the stock mentioned, by selecting a slider between "bullish" or "bearish". This training material is superior to researcher-labeled training material because it reflects unbiased meaning expressed by the author. Alternatively, we apply the maximum entropy (ME) approach and bag-of-word (BOW) approach with Loughran and McDonald (2011) dictionary, and results conclude that the NB yields a higher accuracy and therefore is a better fit for our text sample. To the best of our knowledge, the size of our training sample is the largest among textual analysis studies in accounting/finance area.

of the information content of tweets, which is the mean of the information set of tweets related to a given stock. We show that firm-day twitter information is positively associated with same-day and next-day abnormal returns and Cumulative Abnormal Return (CAR) for up to a week. A long-short hedge portfolio adjusted weekly based on stocks that fall into the most positive decile and most negative decile of average daily StockTwits information earns an average annual abnormal return of approximately 7%. Finally, we also test the validity of users' self-reported sophistication levels. We find that users of StockTwits who describe themselves as "sophisticated" investors: (1) are followed by more users, (2) post more tweets, on average, than other users who identify themselves as "unsophisticated," (3) write longer tweets with more professional financial terms , and (4) write tweets that have a stronger ability in predicting future stock returns.[6] Taken together, the results of all of these validity tests suggests that the new measure does indeed proxy for the degree of information asymmetry between sophisticated and unsophisticated investors.

Using a sample of 6,426 sporadic management forecasts made by public firms during 2009 to 2013, we find that information asymmetry between sophisticated and unsophisticated investors *increase* in the first week following the release of a management forecast, consistent with sophisticated investors interpreting more new information from the public disclosure (e.g. Kim and Verrecchia 1994) than do unsophisticated investors. We also find that from the second week following the management forecast, information asymmetry decreases to a level that is lower than the pre-disclosure level, consistent with management forecasts leveling the informational playing field between sophisticated and

[6] Professional terms are financial and investment terms included in the Dictionary of Financial and Business Terms (University of Toronto). We show that opinion of sophisticated investors has higher coefficients in predicting CAR of the following five trading days than opinion of unsophisticated investors. For detailed classification of investor sophistication, see section 3.2 and Table 6 for detail.

unsophisticated investors, in the longer term. Finally, we find that compared to less precise range forecasts, more precise point forecasts result in less of an increase in information asymmetry in the short term (the first week) but result in a larger long-term reduction in information asymmetry (from the second week following the management forecast). This finding indicates that more precise management forecasts serve to alleviate information asymmetry between sophisticated and unsophisticated investors. More precise forecasts provide less scope for sophisticated investors to trigger relatively more private information, thus limiting the extent to which the public disclosure result in a short-term increase in information asymmetry between sophisticated and unsophisticated investors (e.g. Holthausen and Verrecchia 1990, Lambert et al. 2007).

This study makes a number of contributions. First, we develop, validate, and demonstrate the use of a new measure of information asymmetry between sophisticated and unsophisticated investors. The new measure is superior to existing measures of investors' information in the following ways: 1) It is based upon investors' information that is extracted from an analysis of investors' directly observable tweets (in which investors express their beliefs regarding the investment value of specific stocks) thus it avoids potential identification errors associated with using market aggregates such as bid-ask spread or trade size.[7] 2) The new measure also has distinct advantages compared to using analysts' forecasts as a proxy for investors' information which strictly relies on the

[7] Bid-Ask spread is subject to the impact of transaction cost, and it includes both information asymmetry within sophisticated (or unsophisticated) investors and information asymmetry between the two groups (see Hasbrouck, 1991 for a discussion). Cready et al. (2014) and Frazzini et al. (2012) call into question the validity of using trade size to infer investor sophistications by showing that highly professional investors actively and strategically separate large trading orders into small sized trades, making trade-size a noisy proxy of investor sophistication.

assumption that sell-side analysts can proxy for investors. 3) The new measure does not require a firm to have a large analyst following and is, therefore, more general than analyst-based proxies for investors' information. 4) The new measure can be used to generate data of a far higher frequency (i.e., daily basis) than can analyst-based proxies.

Second, this study shows that: (1) information asymmetry between sophisticated and unsophisticated investors increases in the short-term, i.e., for one week, following public disclosures−specifically, management forecasts; (2) information asymmetry decreases thereafter, i.e., a week after the public disclosure; and (3) more precise public disclosures result in a smaller short-term increase, and a larger long-term decrease in information asymmetry between sophisticated and unsophisticated investors. These results contribute to a number of streams of literature: First, the results provide new evidence that management forecasts increase information asymmetry between sophisticated and unsophisticated investors in the short term but decrease information asymmetry in the long run. Second, the results provide new evidence that sophisticated and unsophisticated investors interpret public disclosures differently in the short-term. Third, the results show that more precise public disclosures serve to "level the informational playing field" between sophisticated and unsophisticated investors.

Finally, this paper contributes to the emerging literature on the role of social media in the capital markets. The finding that the average information contained in StockTwits predicts future abnormal returns indicates that social media data contains useful information that is not fully incorporated by other channels. Additionally, compared to other measures of investors' use of the internet to search for firm-related information, social media provides researchers a lens to observe investor-generated content that presumably reflects their beliefs. Social media analysis, therefore, offers scope for studies of investors' information processing and transmission at an individual investor level.

The remainder of the paper is organized as follows: Section II provides a summary of the relevant literature, the institutional background, and hypothesis development. Section III discusses the construction of the new measure of information asymmetry using StockTwits data, the validity tests of this new measure, and the construction of other financial data. Section IV describes the study design. Section V reports the main empirical results. Section VI concludes the paper.

2. Backgrounds and Hypothesis Development

2.1 Measuring Information Asymmetry between Sophisticated and Unsophisticated Investors

The role of investor sophistication in determining market reactions to financial disclosures has been extensively studied in both the accounting and finance literature. In particular, studies show that sophisticated and unsophisticated investor are asymmetrically informed and that they process public disclosures differently. For example, Wather (1997) and Ayers et al. (2011) show that when forming their earnings expectations, sophisticated investors primarily rely on analyst forecasts, while unsophisticated investors primarily rely on last year's earnings. Studies also show that sophisticated and unsophisticated investors respond differently to public disclosures, resulting in a series of market consequences, including abnormal trading volume around earnings announcements (Utama and Cready 1997), mispricing of accruals (Ali et al. 2000, Collins et al. 2003), and variation in earnings response coefficients (Bartov et al. 2000).

Because researchers cannot directly observe the information sets of sophisticated and unsophisticated investors, existing studies rely on indirect proxies for the degree of information asymmetry derived from aggregate market-level data. In particular, trade size has been widely used as a proxy for the (unobserved) level of sophistication of investors—under the assumption that sophisticated investors are more likely to execute larger trades due to their relatively larger holdings (e.g. Miller 2010, Ayers et al. 2011). However, recent studies (Cready et al. 2014, Frazzini et al. 2012) call into question the validity of these inferences by showing that highly professional investors actively and strategically execute large trades by splitting them into many small trades (with the help of computerized trading system), making trade-size a noisy, and possibly biased, proxy for investor sophistication. Cready et al. 2014 suggest that "such widespread strategic trading behavior by institutional (sophisticated) investors obviously renders trade size a questionable proxy for investor sophistication. "

Social media provides individual investors with a platform to express and exchange their opinions about the investment value of stocks by posting short discussions known as "tweets." These investor-generated statements that reflect investors' beliefs about the investment value of stocks can be used to develop measures of investors' information. Using data from StockTwits.com which contains information for investors' sophistication levels, we textually analyze the tweets of both sophisticated and unsophisticated investors and construct direct measures of the average beliefs of both groups of investors. Using these measures of sophisticated and unsophisticated investors' information, we then develop a direct measure of information asymmetry between sophisticated and unsophisticated investors.

Traditional proxies for information asymmetry (such as bid-ask spread, the probability of informed trading-PIN, etc.) are constructed based upon market-level trading data and are thus subject to potential noise arising from transaction costs (see Hasbrouck, 1991 for a discussion). These measures also capture both information asymmetry between sophisticated and unsophisticated investors (between-group information asymmetry), and the effects of any information heterogeneity within each group (within-group information asymmetry). Using the directly observed tweets of both sophisticated and unsophisticated investors, we generate measures of the information content of both groups and, from these measures of information content; we create a new measure of the information asymmetry between sophisticated and unsophisticated investors that is directly based upon investor-generated statements.

2.2 Public Disclosures and Leveling the Informational Playing Field

Information asymmetry occurs when 1) informed investors possess private information about the firm's value while uninformed investors only have access to public information, or\and 2) investors possess private information of different precision. Information asymmetry is costly to firms as its presence creates an adverse selection problem in the stock market. When investors with superior information trade on the basis of their private information, investors at an information disadvantage will seek price-protect, thus increases firms' cost of capital (e.g. Grossman and Stiglitz 1980, Kyle 1989; Lambert and Verrecchia 2010). Information asymmetry between sophisticated and unsophisticated investors is also of concern to regulators as it creates an uneven informational playing field, potentially compromising market fairness (see Lev 1988).

While it seems intuitive that public disclosures should reduce information asymmetry by making information privately possessed by some investors publicly available to all investors, theories suggest that the real impact of public disclosures on information asymmetry may not necessarily be so benign. Public disclosures can affect the degree of information asymmetry between investors of varying sophistication levels through two channels. First, in the pre-disclosure period sophisticated investors may have access to private information that is not available to unsophisticated investors. In this case, public disclosures will decrease information asymmetry between heterogeneously informed investors by making private information publicly available to all investors (see Kim and Verrecchia 1991b). In other words, when public disclosures are released, investors of various sophistication levels adjust their information sets towards the public signal, thus increasing the commonality of traders' information and decrease the information asymmetry between sophisticated and unsophisticated investors. Second, public disclosures can potentially cause investors with different information processing skills to generate new private information, resulting in an increase of information asymmetry between sophisticated and unsophisticated investors (Kim and Verrecchia 1994 and 1997, and Indjejikian 1991).

Earnings announcements have been widely used as a setting to study changes in investors' information around public disclosures. Earnings announcements provide significant amounts of information regarding the firm value and are therefore likely to affect investors' information. For example, Lee et al. (1993) show that bid-ask spreads increase around earnings announcements; they argue that earnings announcements trigger private information search activities, resulting in an increase in bid-ask spreads. Also, Yohn (1998) shows that bid-ask spreads increase in the four days before earnings announcements remain high for earnings announcement days and the days immediately following, but bid-ask spreads one week after the earnings announcements are not significantly different from the pre-announcement level.

However, these results for earnings announcements may not generalize to other types of public disclosures, such as management forecasts. First, earnings announcements are regular and anticipated disclosures. Theory suggests (e.g. Kim and Verrecchia 1991b) that investors specifically collect private information prior to anticipate disclosures, known as the anticipation effect. For earnings announcements, it is hard to disentangle the effects of pre-announcements information acquisition activity from the effects of announcements per se. Also, sophisticated investors' ability to better interpret the announcements – to trigger new private information, as discussed in Kim and Verrecchia 1994 – might be because they acquire some private information in anticipation of the announcements. Empirical evidence suggests that the anticipation of earnings announcements alone increases information asymmetry as traders search for private information in an attempt to profit from the upcoming public disclosure (e.g. Yohn 1998).

Second, earnings announcements mainly report confirmatory historical performance data (i.e. income statement) and historical values (i.e. balance sheet). Although this information is useful in predicting firms' future performance, it demands a higher level of knowledge and skills for investors to interpret to form beliefs about firms' future performance correctly. Management forecasts, on the other hand, provide more precise estimates of firms' future performance (usually EPS numbers) and require fewer interpretation skills (as EPS numbers are usually provided). As a result of such fundamental difference between the two types of announcements, the findings of Yohn (1998) that bid-ask spreads do not change significantly in the long term following earnings announcements may not hold for the case of management forecasts.

This study uses sporadic management forecasts as the setting to study the impact of public disclosure on the information asymmetry between sophisticated and unsophisticated investors. Sporadic management forecasts are usually unanticipated by investors, therefore minimizing the impact of investors searching for private information ahead of the disclosure. In addition, management forecasts offer a cleaner setting to identify the level of precision of disclosures (Baginski et al. 1993), allowing for empirical tests of how variation in the precision of public disclosures determines the effect of public disclosures on the degree of information asymmetry between sophisticated and unsophisticated investors.

Ex-ante, it is unclear whether management forecasts would decrease or increase the information asymmetry between sophisticated and unsophisticated investors. On the one hand, if management forecasts reveal to unsophisticated investors information that was previously only privately possessed by sophisticated investors, then management forecasts would *reduce* the degree of information asymmetry between the two classes of investors. On the other hand, if the superior information processing skills of sophisticated investors allow them to extract more private information from a management forecast than unsophisticated investors, then management forecasts could potentially *increase* the information asymmetry between sophisticated and unsophisticated investors. The first hypothesis stated in the null form:

H1: Information asymmetry between sophisticated and unsophisticated investors does not change around the disclosure of sporadic management forecasts.

Next, we examine whether the impact of public disclosures on information asymmetry between sophisticated and unsophisticated investors varies with the level of precision of the public disclosure. We expect the impact of management forecasts on the degree of information asymmetry between sophisticated and unsophisticated investors to vary with the level of precision of the management forecast (point vs. range forecast) because: First, studies suggest that diversely informed investors update their prevailing beliefs to incorporate the public disclosures, and that their belief update is associated with the precision of the public disclosure (Kim and Verrecchia 1991). Second, studies show that investors interpret public disclosures differently (i.e. differential interpretation of disclosure).[8] As a result, a less (more) precise public disclosure may provide traders with superior information processing skills more (less) scope to trigger new private information, as the potential expected payoff from such processing of the public disclosure increased (decreased). Following prior studies (e.g. Baginski et al. 1993, Pownall et al. 1993, Hirst et al. 2008), we use point management forecast as the proxy for more precise management forecast, and range forecast as the proxy for less precise management forecast. The second hypothesis, stated in the null form is:

H2: Point and range management forecasts are associated with the same change in information asymmetry between sophisticated and unsophisticated investors.

3. Data and Sample

[8] Kandel and Pearson (1995) and Kim and Verrecchia (1997) show that announcements themselves convey different things to investors with different sophistication levels. Using analyst forecast data, Barron, Byard and Kim (2002) shows that analysts extract or develop private information from the public disclosures.

3.1 Construction of Measure of Information Asymmetry between Sophisticated and Unsophisticated Investors using Social Media Data

Social media provides people with a new platform to express and exchange their opinions and ideas with a large number of peers. As such, the rapid development of social media networks in recent years can be expected to affect the origination and transmission of information to capital markets. Indeed, some pioneering studies examine investment-related social media networks and provide evidence suggesting that the information as expressed on such social media networks has a significant impact on the capital market.[9]

Stocktwits.com is the most popular investment-dedicated social media micro-blogging website in the US. As of July 2014, StockTwits has more than 600,000 active users and an average monthly posting volume of more than 3 million tweets. Per StockTwits, the posts "are viewed by an audience of over 40 million across the financial web and social media platforms."[10] Stocktwits.com features a platform on which users can post a short paragraph of no more than 140 characters (called a tweet) that relates to the investment potential of one or more specific stocks. These tweets are then posted to the main board of the website as well as specific sub-pages sorted by stock tickers and can be viewed by other users.

StockTwits.com provides an ideal setting for measuring information asymmetry between sophisticated and unsophisticated investors for the following reasons: First,

[9] For example, a recent study in computational science by Bollen et al. (2011) shows that the content of Twitter.com predicts future stock returns. More recently, Chen, De, Hu, and Hwang (2014) suggest that the information content of Seeking Alpha, a popular investment related blog website, can predict future stock price performance.

[10] From introduction of StockTwits.com (http://stocktwits.com/about).

StockTwits.com is a venue dedicated to of investment potential of stocks. Tweets posted on this website are investment-focused, thus minimizing potential inclusion of noisy or tweets that are not investment-related (e.g., consumer-related comments that focus on firms' produces and services). Second, investors who post on StockTwits self-report their sophistication levels as "Novice, Intermediate or Professional," allowing for identification of information asymmetry between sophisticated (professional) and unsophisticated (novice) investors. Moreover, while self-reported, the StockTwits data also allows for tests of the validity of these self-reported investor sophistication levels (see Section 3.2 for detail). Third, StockTwits.com features a "Cash-tagging" design to clearly identify the ticker each tweet refers to, thus enabling the development of programs to clearly extract the company each post refers to, minimizing the potential for the misclassification of firms which is common among other internet-based data sources. [11]

In this study, we extracted and textually analyzed 11 million tweets posted between July 2009 and December 2013 by around 290,000 users.[12] First, we carefully remove tweets (and re-tweets) posted using the official accounts of news media organizations such as NYTimes and WSJ and posts originating from the official StockTwits accounts of firms.

3.2 Naïve-Bayes Classification Approach

Naïve-Bayes is a machine-learning classification method that assign text to its most likely category base on a probabilistic relationship between features (word, word groups)

[11] Each tweets on StockTwits is tagged with the ticker symbol (expressed as $GOOG or $AAPL) that the author is referring to, in practice called "Cash-tagging". This special design provides a mechanism to clearly extract the company reference in each post with no misclassification. By searching for the "$", program can automatically extract the ticker each tweet is related to.

[12] We want to thank Chris Corriveau at StockTwits.com for granting me access to StockTwits data and API Service.

and the category (positive, neutral, negative) the algorithm has learned from a given training set. Formally, this approach assign a sentence s, containing n words, $\{w_1, w_{2,} \ldots, w_n\}$, to one of m categories $c^* \in \{c_1, c_{2,} \ldots, c_m\}$, by maximizing the conditional probability $Prob(c|s)$ that the sentence belongs a certain category (positive, neutral or negative):

$$c^* = argmax\ Prob(c|s).$$

Applying Bayes' theorem with the "naive" assumption of independence between every pair of features $\{w_1, w_{2,} \ldots, w_n\}$, the classification rule can be stated as follow:

$$c^* = argmax\ Prob(c) \prod_{j=1}^{m} prob(w_n|\, c)$$

Compared to bag-of-word method that relies on a pre-classified dictionary to determine category based on the appearance of certain words, Naïve-Bayes classification method has a higher accuracy by adapting to the words that appear in a specific domain and their probabilistic relation to a certain opinion category. Despite their apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many real-world situations.

3.3 Implementation of Naïve-Bayes Classification Approach

The quality of training material is of vital importance to the success of Naïve-Bayes classification method; high-quality training material should provide sufficient feature sets and similar language domain as the material to be classified. StockTwits has a function called "sentiment selector": when a user post a tweet on to StockTwits, she/he has the option to choose between "positive" or "negative" by clicking on a slider next to the textbox. This particular feature of StockTwits offers us a naturally classified training material for our Naïve-Bayes algorithm. In total, we have a total of about 1.1 million pre-classified tweets as our training material. Compared to other studies that use hand-classified sentences as training material, this setting helped us avoid the subjectivity of hand classification. Also, to the best of our knowledge, the size of our training set is largest among textual analysis studies in accounting/finance field.

After training our Naïve-Bayes classifier with the training set we acquire from StockTwits, we let the "trained" algorithm classify the remaining sample of 10 million tweets. Because the context of social media might be different from other known context such as 10-K MD&A or news articles, it is unknown whether Naïve-Bayes might best approach to classify social media text. To compare the accuracy of various text classification approaches in social media context, following Li (2010) we estimate in-sample and out-of-sample accuracy for machine learning methods including Naïve-Bayes (NB), Maximum Entropy (ME) and Support Vector Machine (SVM) methods, and dictionary-based method using Laughran and McDonald (2011) financial dictionary. The result is reported in Panel A of Table 2.

The result reported in Panel A of Table 2 shows that Naïve-Bayes classification method provides in-sample accuracy of 89.15% and out-of-sample accuracy of 86.41%, and an average accuracy of about 88%. To our best knowledge, the classification accuracy we have achived is highest among textual analysis studies in accounting/finance literature, we attribute this relatively high accuracy rate to two factors: large size of our training sample; larger proportion of meaningful words in a tweet due to 140-character limitation in a tweet motivating users to reduce the use of meaningless words. From the result, we also conclude that Naïve-Bayes has the best performance among various machine learning approaches and the traditional dictionary based approach in the context of social media text.

To estimate the information asymmetry between sophisticated and unsophisticated investors, we focus on the tweets posted by "Novice" and "Professional" users.[13] We first use textual analysis to extract the information (positive, neutral or negative) of each tweet. Then, based upon users self-reported sophistication levels, we calculate the mean information sets of sophisticated and unsophisticated investors for a given stock. We then construct my new measure of information asymmetry between sophisticated and unsophisticated investors as the absolute value of the difference between sophisticated and unsophisticated investors' information sets.[14] Because we use financial data of NYSE and NASDAQ listed firms, following Antweiler and Frank (2004), we align tweets posted within the trading hours of 9:30 am to 4:00 pm to the same trading day and tweets made after hour (and pre-market) are matched with the market level data of the next trading day. Because these tweets can only have an effect on the market indicators of the next trading day.

3.4 Validity Tests of Social Media Based Information Asymmetry Measures

Before proceeding to empirical test, we first undertake tests to assess the validity of this new measure of information asymmetry between sophisticated and unsophisticated investors. Panel B of Table 2 shows that variation in the measurement of information is

[13] In section 3.2 We provide validity test of the self-reported sophistication levels, and In Table 6, we provide result of these validity tests.

[14] We require at least five tweets posted by identical sophisticated users and unsophisticated users on the same day to calculate the information asymmetry between these two classes of investors. The measures used in the validity tests are mean opinion of all investors, and information asymmetry measure used in the validity tests is the standard deviation of all tweets by all investors.

positively correlated (all significant at p=0.00 level) with existing proxies for opinion divergence such as raw and abnormal return (Miller 1977), and raw and abnormal trading volume (Bamber, Barron and Stevens 2011).

We first examine the relationship between firm-level aggregated twitter information and stock performance. In Table 3, we show that aggregated firm-day twitter information is positively correlated with contemporaneous abnormal returns, abnormal returns of the next trading day, and Cumulative Abnormal Returns (CAR) for up to five trading days.[15] A hedged portfolio holding a long position of the most positive decile and a short position of the most negative decile of firm-day Twitter information earns an average abnormal return of 7% annually. These results are robust to the exclusion of trading days around earnings announcements.[16] As reported in Table 4, after controlling for firm characteristics that may contribute to the cross-sectional variation in market performances, firm-day StockTwits information is still significantly positively (p-value<0.01) abnormal returns for the next trading day, and the cumulative abnormal returns for the following three trading days. Similarly, as reported in Table 5, the divergence of opinions among investors following the same firm, as measured by the standard deviation of information scores for the individual tweets for a given firm on a given day, has a positive and significant coefficient in predicting abnormal trading volume for up to five trading days.

[15] We Estimate the CAR as buy and hold excessive return of a given firm over the benchmark firm matched on Size and Book to Market (5x5 groups). The formation of the benchmark portfolio follows instruction on Kenneth French's personal website.

[16] Studies show that unsophisticated investors may potentially contribute to post-earnings announcements drift (e.g. Hirshleifer et al. 2008). To eliminate the possible impact of PEAD, We exclude the trading days around earnings announcements to show the robustness to the exclusion of earnings announcements events.

Next, we examine the validity of user self-reported sophistication levels. We expect users with different sophistication levels to have different information about the investment potential of stocks. As a result, the twitting behaviors should vary across sophistication levels, so that tweets posted by users of differing levels of sophistication should differ in their abilities in predicting future returns. In Table 6A, we show that in my sample, 55% of investors identify themselves as "unsophisticated" and about 14% identify themselves as "sophisticated." Investors who identify themselves as "sophisticated" are followed by more users (103 vs. 12 followers, on average) and, on average, sophisticated investors post more tweets (294 vs. 85 tweets) per year than investors who identify themselves as "unsophisticated." Tweets by "sophisticated" investors are significantly longer and contains more professional terms.[17] These differences in the tweeting behavior of self-reported sophisticated and unsophisticated investors are all significantly different at conventional levels (i.e., t-tests for differences in means are all significant at the 1% level). Additionally, as reported in Table 6B, tweets by sophisticated investors have higher coefficients in predicting future abnormal return (coefficients of 0.0375 vs. 0.0252, and a p-value of the F-test=0.00) than tweets by unsophisticated investors. In summary, we find consistent evidence that investors' self-reported sophistication levels capture real differences in investors' underlying abilities and skills for collecting and process information related to the investment potential of stocks.

3.5 Management Forecast Sample and Other Financial Data

[17] Professional terms are financial and investment terms included in the Dictionary of Financial and Business Terms (University of Toronto).

Using the Thomson Reuters First Call database, we downloaded 23,166 management forecasts sample issued between 2009 and 2013. We eliminate “bundled” management forecasts made within -3 to +3 days of earnings announcements, identified using earnings announcement dates from Compustat and I/B/E/S.[18] Following the identification method of Tang (2012), we further eliminate regularly made management forecasts.[19] This data selection procedure yields a sample of 6,425 sporadic management forecasts. Following Baginki et al. (1993), we identify point and range forecasts conditioning on whether a single value or a range value has been released, and the explanatory phrases provided by First Call. Management forecasts made with only one estimation, and that do not contain conditioning phrases such as “greater than”, “less than” or “no more than”, are labeled as point forecasts; management forecasts with both upper and lower bonds, or with one bound and conditioning phrases are labeled as range forecasts (e.g. “EPS is greater than $1”). This identification method yields a total of 1,061 point forecasts (16.72%), and 5,364 range forecasts (83.28%). The detailed decomposition of the management forecasts data is reported in Table 1 Panel D.

The original tweets sample spans from July of 2009 to end of 2013, covering 8,721 identical tickers. After removing 1) Tickers that are non-US, 2) financial assets other than stocks (such as futures for indexes and exchange rates), we matched the ticker with Compustat and I/B/E/S to get financial and analyst following data. Our final sample

[18] Following prior studies (e.g. DellaVigna and Pollett 2009), when the EA dates are available in both databases and are difference from each other, We take the date that report the earlier of the two.

[19] We applied a similar selection method of Tang (2014) to eliminate the recurring and therefore predictable management forecast; We look at two consecutive years and in which week the sample forecast happened, if it happens in the same week of year t and year t-1, then both forecasts will be dropped from our sample.

consists of 19,173 firm-years observations (for 4,701 firms). Stock price and trading data are from CRSP. The detailed construction of variables is reported in Appendix A.

4. Research Design

To examine the effect of management forecasts on the degree of information asymmetry between sophisticated and unsophisticated investors, we calculate daily measures for the degree of information asymmetry for one week before a sporadic management forecast is issued, to one week after the forecast; we also calculate daily information asymmetry for longer windows after the management forecast is issued. We compare the level of information asymmetry in the post management forecast periods with the level of information asymmetry in the pre-forecast period. Following Rogers et al. 2009, we calculate the change in information asymmetry using the following formula:

$$Change_{Week1} = ln\left(\frac{\sum_{1}^{7} Info\ Asymmetry}{\sum_{(-7)}^{(-1)} Info\ Asymmetry}\right) \qquad Change_{Week2} = ln\left(\frac{\sum_{8}^{14} Info\ Asymmetry}{\sum_{(-7)}^{(-1)} Info\ Asymmetry}\right)$$

$$Change_{Week3} = ln\left(\frac{\sum_{15}^{21} Info\ Asymmetry}{\sum_{(-7)}^{(-1)} Info\ Asymmetry}\right) \qquad Change_{Week4} = ln\left(\frac{\sum_{22}^{28} Info\ Asymmetry}{\sum_{(-7)}^{(-1)} Info\ Asymmetry}\right)$$

Using a weekly measure of information asymmetry (aggregation of daily information asymmetry) eliminates the day-of-the-week effect on information asymmetry arising from factors such as day-of-the-week variation in investors' attention, market liquidity, etc. Because measures of change in information asymmetry are calculated relative to the same benchmark—level of information asymmetry one week prior to forecasts—they should be interpreted as the net effect of management forecasts on the information asymmetry between sophisticated and unsophisticated investors. To further control for other firm and market characteristics that may affect information asymmetry, we employ regression analysis using the following equation:

$$\begin{aligned} &Change_{Week(i,t)} = \alpha_0 + \beta_1 Management\ Forecast_{i,t} + \beta_2 Size_{i,t} + \\ &\beta_3 Book\ to\ Market_{i,t} + \ \beta_4\ Leverage_{i,t} + \ \beta_5 Analyst\ Following_{i,t} + \\ &\beta_5\ |Forecast\ Error|_{i,t} + \beta_6 News_{i,t} + \beta_7\ Profitbility_{i,t} + \ \beta_8\ Industry_{i,t} + \varepsilon_{i,t} \qquad (1) \end{aligned}$$

$$\begin{aligned} &Change_{Week(i,t)} = \alpha_0 + \beta_1 Point\ Forecast_{i,t} + \beta_2 Size_{i,t} + \ \beta_3 Book\ to\ Market_{i,t} + \\ &\beta_4\ Leverage_{i,t} + \ \beta_5 Analyst\ Following_{i,t} + \beta_5\ |Forecast\ Error|_{i,t} + \beta_6 News_{i,t} + \\ &\beta_7\ Profitbility_{i,t} + \ \beta_8\ Industry_{i,t} + \varepsilon_{i,t} \qquad (2) \end{aligned}$$

Equation 1 is used to test hypothesis 1. In Equation 1, *Management Forecast* is a dummy variable that equals one if firm *i* issued a sporadic management forecast on day *t*, and 0 if firm *i* issued no management forecast on day *t*. When firm *i* has no management forecast on a given day *t*, we expect that the dependent variable that captures the change of information asymmetry in the week following day t relative to the preceding day *t* to be zero. When a firm issues a management forecast on day *t*, as discussed in the hypothesis development section, $\beta_1$would be significantly different from zero, indicating the impact of management forecast on information asymmetry. A positive $\beta_1$indicates an increase in information asymmetry following a management forecast, while a negative β_1 indicates a decrease in the information asymmetry following the disclosure of a management forecast.

Equation 2 is used to test hypothesis 2. In equation 2, Point Forecast is a dummy variable that equals to 1 if the forecast is a point forecast, and 0 if the forecast is a range forecast. The sample used in this regression is limited to management forecast sample only. Therefore results should be interpreted as the difference in the impact of a point forecast on information asymmetry and the impact of range forecast on information asymmetry. As discussed in hypothesis 2, a positive β_1 indicates that a point forecast is associated with more information asymmetry than a range forecast, while a negative β_1 indicates that a point forecast is associated with less information asymmetry than a range forecast. This regression is estimated in short term (one week) and longer term (up to four weeks) to study how the impact varies across time. We expect β_1 to be negative or not significant in short term, and negative in the longer term.

The control variables are from the most recent annual filings prior to the issue date of the management forecast. Specifically, firm characteristics such as size and market-to-book ratios are calculated using data from firms' most recent 10-K available from Compustat. Analyst following data is the natural log of 1 plus the number of analysts following the firm. The first regression includes firm fixed effect; the second regression includes industry fixed effect; standard errors are clustered by industry (2-digit SIC).

5. Empirical Results

5.1 Management Forecasts and Change in Information Asymmetry (H1)

5.1.1 Univariate Result

Figure 1 plots the average daily information asymmetry between sophisticated and unsophisticated investors relative to day 0, the issuance date for a sporadic management forecast. On average, information asymmetry between sophisticated and unsophisticated investors does not average to zero because sophisticated and unsophisticated investors each possess different information sets, even during the periods when there is no management forecast disclosure. We can identify a few patterns from Figure 1: First, information asymmetry between sophisticated and unsophisticated investors does not increase significantly in the three-week window prior to the issuance of sporadic management forecasts, consistent with little additional private information collecting activities prior to sporadic management forecasts, consistent with sporadic management forecast not being anticipated by investors. Second, information asymmetry spikes on the day of management forecast issue date and stays higher than the pre-disclosure level for the first week following the forecast. This is consistent with new private information being triggered by management forecast, thus increasing information symmetry (e.g. Kim and Verrecchia 1994), at least in the short term. Third, after roughly a week following the management forecast issue data information asymmetry decreases to a level that is lower than the pre-disclosure level, and information asymmetry continues to decrease in the third week following the management forecast issue data. This finding shows that management forecasts increase information asymmetry in the short term, but decrease it in the longer term, consistent with management forecast leveling the informational playing field between sophisticated and unsophisticated investors in the longer term, i.e., after one week.

Table 7 Panel A reports a univariate test of changes in information asymmetry around sporadic management forecasts for MF-firms vs. non-MF firms (firms that did not issue MFs in the -30 to +30 window) matched Size, Book to Market and 2-digit SIC code. For non-MF firms, day 0 is set to the same day that its matched MF firm issues a management forecast. For matched non-MF firms, the changes in information asymmetry over the four weeks' periods relative to information asymmetry in the preceding week are not significantly different from zero. However, MF firms experience higher information asymmetry in the first week (significant at 1% level), lower information asymmetry in the second and third weeks (significant at 5% level), and information asymmetry in the fourth week continues to be lower (marginal significant at 10% level) than the pre-disclosure level.

5.1.2 Multivariate Result

Table 8 reports regression results for equation 1 which examines the change in information asymmetry around MFs, controlling for firm and market characteristics that are likely to contribute to these changes. In Column 1 of Table 8, the dummy variable MF has a significantly positive coefficient ($p<0.01$, two-tailed), suggesting that information asymmetry between sophisticated and unsophisticated investors increase in the short term following sporadic management forecasts. However, in Columns 2 and 3 of Table 8, the coefficients on *Management Forecast* are negative and significant at the 5% level, suggesting that in the longer term following sporadic management forecasts, information asymmetry between sophisticated and unsophisticated investors decrease to a level that is lower than that in the period prior to the management forecast issue data.

5.2 Point Forecasts and Range Forecasts Subsamples (H2)

5.2.1 Univariate Result

Next, to explore the effect of variation in the precisions of public disclosures on the information asymmetry between sophisticated and unsophisticated investors we partition the management forecast sample into point and range forecasts. Figure 2 separately plots the mean daily information asymmetry between sophisticated and unsophisticated investors relative to day 0 for point and range forecasts. We can conclude the following from this figure: First, pre-disclosure information asymmetry is not significantly different between these two types of management forecasts. Second, on average point forecasts have a smaller short-term increase in information asymmetry, consistent with more precise public disclosure triggering relatively less new private information, at least in the short term. Third, post-disclosure information asymmetry is lower for point forecast vs. range forecast, indicating that more precise public disclosure results in a larger long-term reduction in information asymmetry between sophisticated and unsophisticated investors. For range forecasts, the long-term post-disclosure level of information asymmetry is unchanged from that in the pre-disclosure period. However, for point forecasts it is lower. The long run reduction in information asymmetry following management forecasts is therefore attributable to the minority of management forecasts that are point forecasts. Table 7 Panel B reports a univariate test of the mean changes in information asymmetry around point and range forecasts. Point forecasts are associated with less of a short-term increase, and more long-term decrease, in information asymmetry than is the case for range forecasts.

5.2.2 Multivariate Result

Table 9 reports regression results for equation two which examines the change in information asymmetry around point vs. range management forecasts. The testing sample is limited to days when management forecasts are issued, so the results should be interpreted as the difference in the impact of a point vs. range forecasts on information asymmetry between sophisticated and unsophisticated investors. In Column 1 of Table 9, the dummy variable Point has a positive coefficient that is not significant, suggesting that, after controlling for firm and market characteristics, the short-term change in information asymmetry between sophisticated and unsophisticated investors is not significantly different between point and range forecasts. However, in Column 2 of Table 9, the coefficient is negative and significant at the 5% level, suggesting that in the longer term information asymmetry between sophisticated and unsophisticated investors decrease more for point forecasts than for range forecasts. These findings are consistent with hypothesis

2 that point forecasts reduce information asymmetry between sophisticated and unsophisticated investors more than range forecasts.

6. Conclusions

This paper examines how public disclosures affect information asymmetry between sophisticated and unsophisticated investors using the setting of sporadic management forecasts. The level of information asymmetry between sophisticated and unsophisticated investors is specifically important to firms because it increases firms' cost of capital through adverse selection problem. It is also important to regulators because its existence compromises the fairness of the stock market by making the informational playing field unleveled for sophisticated and unsophisticated investors. Despite its importance, there is no existing direct measure of the degree of information asymmetry between sophisticated and unsophisticated investors.

In this study, we use textual analysis to develop, demonstrate, and validate a direct measure of information asymmetry between sophisticated and unsophisticated investors based upon 11 million tweets posted on StockTwits.com, the most popular investment-related social media micro-blogging website in the US. Using this new measure, we find that information asymmetry between sophisticated and unsophisticated investors increase in the short term following the issuance of sporadic management forecasts, and decrease to longer term to a level that is lower than that in prior to the issuance of the management forecast. Additionally, we document that point forecasts reduce information asymmetry between sophisticated and unsophisticated investors more than range forecasts.

This paper makes a number of contributions to the literature. First, it provides a new and direct measure of information asymmetry between specific classes of investors. This new measure overcomes the potential identification errors when assessing the information set of investors using analyst forecasts, and moreover, it is available at a higher frequency and applicable in a wide range of studies. Second, this paper adds to the literature on effects of public disclosures on the information asymmetry by showing that information asymmetry could arise from public disclosures due to differential interpretations of the disclosure, and that more precise disclosures have a greater effect in reducing the information asymmetry between sophisticated and unsophisticated investors. Finally, this paper contributes to prior studies on investor sophistication by providing evidence that sophisticated investors and unsophisticated investors interpret public disclosures differently, at least in the short term.

Taken together, this paper suggests that when studying the market consequences of public disclosures, one must take into consideration that these market consequences may be different for sophisticated vs. unsophisticated investors. It also suggests that some types of the disclosure may perform better than others, in helping unsophisticated investors compete against sophisticated investors.

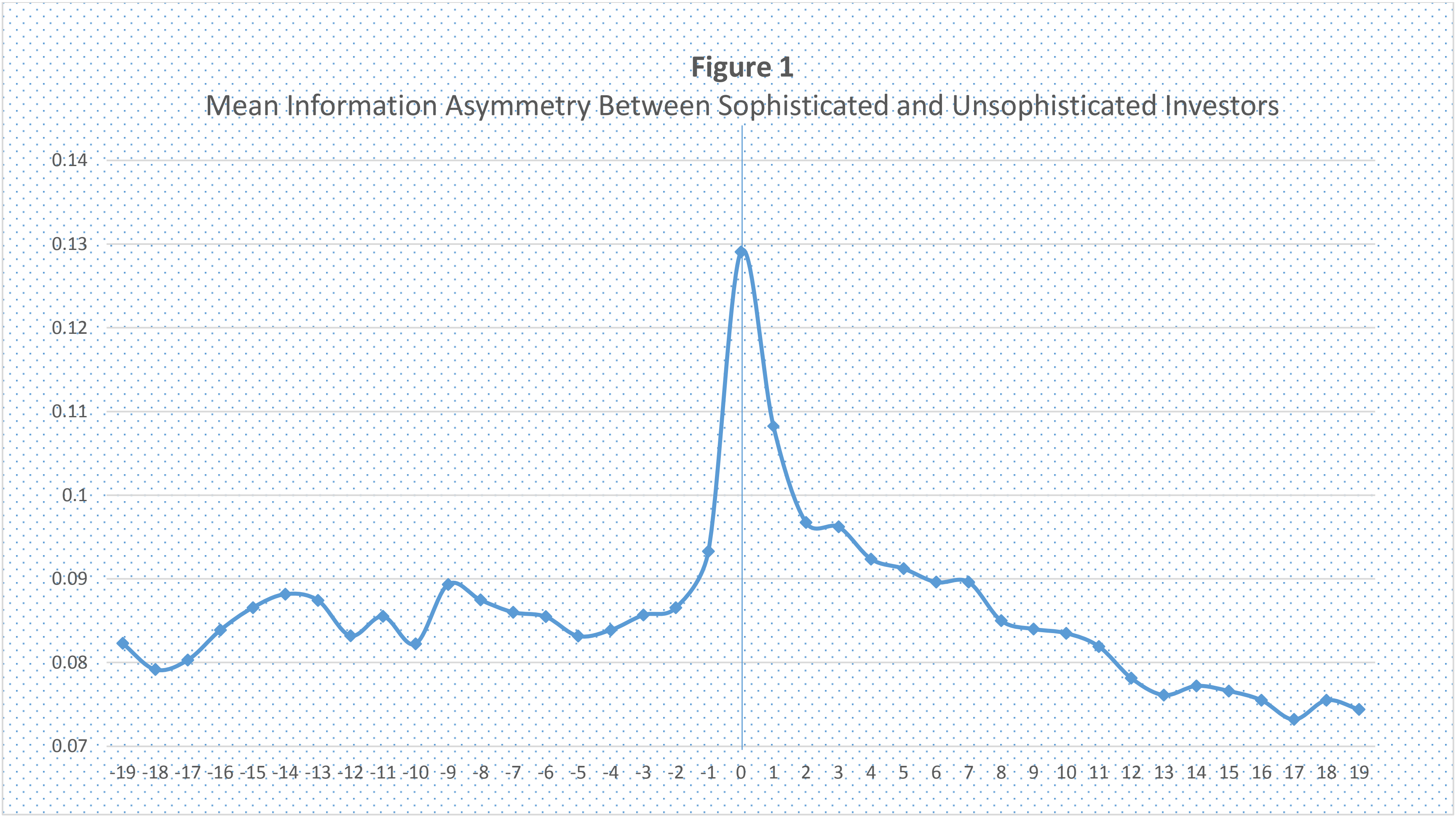


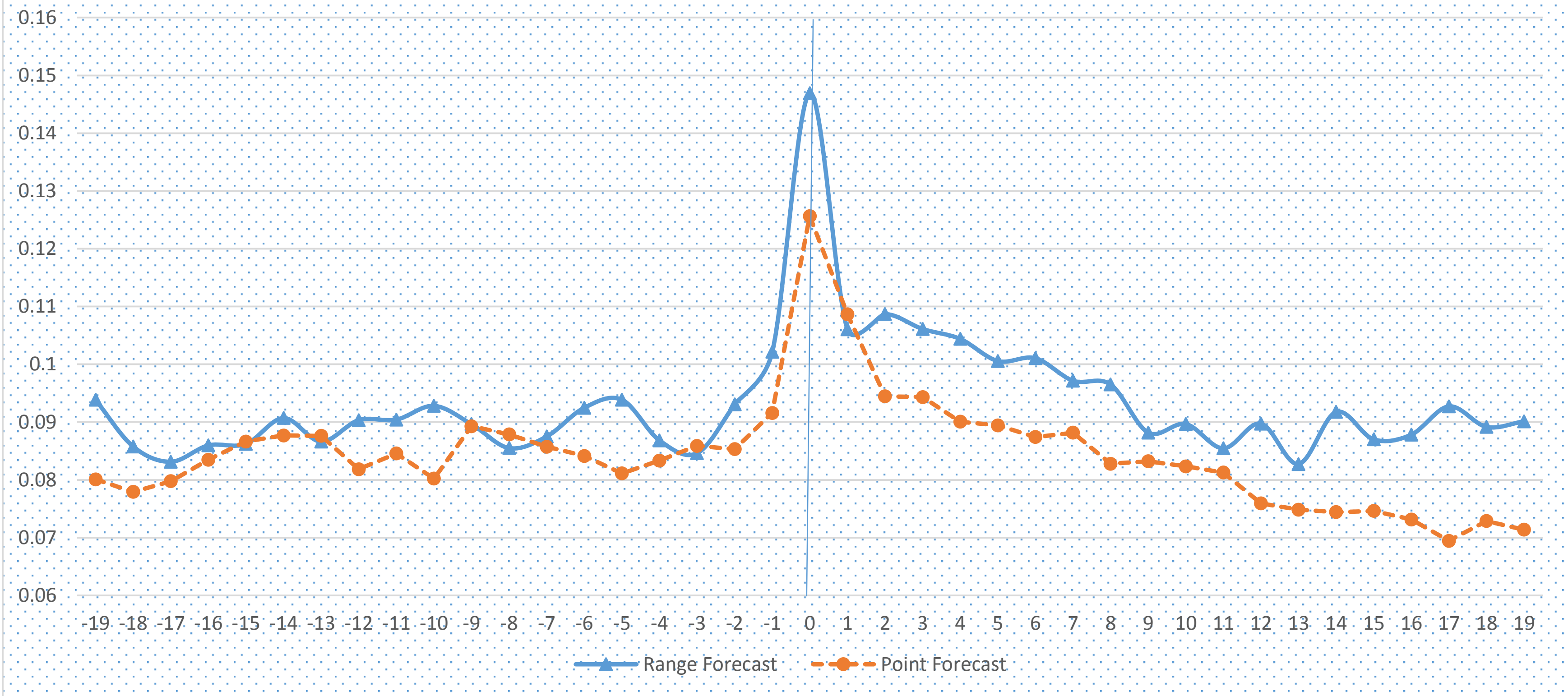
Figure 1

Mean Information Asymmetry Between Sophisticated and Unsophisticated Investors

This chart plots the mean daily information asymmetry between sophisticated and unsophisticated investors. Day 0 is the day that sporadic management forecast is issued by a firm, day t represents the information asymmetry on a trading day post (pre) the disclosure date. Information Asymmetry is the distance between sophisticated investors' information and unsophisticated investors' information, based on Tweets posted on StockTwits.com.

Figure 2

Mean Information Asymmetrey Between Sophisticated and Unsophisticated Investors Around Sporadic and Range Management Forecasts



This chart plots the mean daily information asymmetry between sophisticated and unsophisticated investors around 1) Point Forecast 2) Range Forecast. Day 0 is the day that sporadic management forecast is issued by a firm, day t represents the information asymmetry on a trading day post (pre) the disclosure date. Information Asymmetry is the distance between sophisticated investors' information and unsophisticated investors' information, based on Tweets posted on StockTwits.com.

Table 1 Descriptive Statistics

Panel A: Summary Statistics of Market Data

Variable	N	Mean	SD	P25	Median	P75
Firm Daily Information	1,440,613	-0.326	0.694	-0.693	-0.405	0.173
Daily Raw Return	660,480	0.001	0.034	-0.012	0.000	0.012
Daily Abnormal Return	660,480	0	0.032	-0.010	-0.001	0.009
CAR(1,3)	660,480	0	0.052	-0.019	-0.001	0.017
CAR(1,5)	660,480	0	0.065	-0.024	-0.001	0.022
S&P 500 Index Return	660,480	0.001	0.009	-0.004	0.001	0.005
News Index	386,735	0.651	0.301	0.000	0.692	1.609

Panel B: Summary Statistics of Firm Characteristic Data

Variable	N	Mean	SD	P25	Median	P75
Book to Market	19,173	0.864	1.831	0.324	0.577	0.956
Size	19,173	12.725	2.097	11.191	12.650	14.130
Leverage	20,133	0.223	0.238	0.028	0.166	0.346
Dividend Payout	20,145	0.013	0.026	0.000	0.000	0.015
Profitability	19,507	0.065	0.243	0.027	0.096	0.154
R&D Intensity	19,999	0.121	0.490	0.000	0.000	0.043
Sales Growth	16,105	-0.063	0.596	-0.169	0.016	0.188
Number of Analyst	20,252	9.455	8.732	3.000	7.000	13.000
Analyst Following	20,252	2.019	0.828	1.386	2.079	2.639

Panel C: Summary of Twitter Characteristics

Variable	N	Mean	SD	P25	Median	P75
length	17,857,953	15.06	7.11	9.00	15.00	21.00
Positive Words (LM)	17,857,953	0.22	0.53	0.00	0.00	0.00
Negative Words (LM)	17,857,953	0.54	0.86	0.00	0.00	1.00
Meaningful Words	17,857,953	0.76	1.04	0.00	0.00	1.00

Panel D: Management Forecast Characteristics

Year	No. of Management Forecasts	No. of Sporadic Forecasts	Percentage	Mean Range Width	Point Forecast	Percentage of Point/Sporadic
2009	4,720	1,362	28.86%	0.00759	230	16.89%
2010	5,028	1,451	28.86%	0.00367	204	14.06%
2011	4,777	1,325	27.74%	0.00401	204	15.40%
2012	4,943	1,331	26.93%	0.00525	237	17.81%
2013	3,698	957	25.88%	0.00322	186	19.44%
Total/Mean	23,166	6,426	27.65%	0.00475	1061	16.72%

Note: The variables are defined in appendix A.

Table 2 Validity Check of Information Asymmetry Measure

Panel A: Naïve Bayes Learning Approach versus other Machine Learning Approach and General Dictionary Approach by Laughran and McDonald (2011)

	Accuracy (%)	False Positive (%)	False Negative (%)	False Neutral (%)
Naïve-Bayes (in-sample validation)	89.15	3.76	1.77	5.32
Naïve-Bayes (out of sample 10-fold validation)	86.41	3.88	2.43	7.28
Support Vector Machine	79.17	6.31	4.00	10.52
Maximum Entropy	76.44	7.98	5.21	10.37
Financial Dictionary (LM 2011)	61.22	4.77	10.35	23.68

Panel B: Information and Divergence of Information at Firm-Day Level

	1	2	3	4	5	6	7
	Firm-Day Info	Divergence of Info	Raw Return	Abnormal Return	Trading Volume	Abnormal Volume	Abnormal Volume of T+1
1	1						
2	-0.1076*	1					
3	**0.0296***	**0.0617***	1				
4	**0.0272***	**0.0583***	0.9642*	1			
5	-0.0273*	**0.1488***	0.0041*	0.0073*	1		
6	0.0002	**0.0021***	0.2630*	0.2745*	0.0034*	1	
7	0.0001	**0.0014***	0.1053*	0.1113*	0.0084*	0.0054*	1

This table presents some basic characteristics of text-based information extraction. Panel A presents the summary of text-based information. Approximately 10% of all tweets come with a user-identified Information indicating “bullish” or “bearish”, rest are classified as “neutral”. Panel B presents the correlation between user-identified information, reader-identified information and text-based information. Panel C presents the correlation between aggregated firm-day information and variation of Information (standard deviation of information for tweets), and several market level measures.

* indicates significant at 0.1 level ** indicates significant at 0.05 level *** indicates significant at 0.01 level

Table 3: Return of Portfolio Formed on Firm-Level StockTwits Information Content

		Most Negative								Most Positive				
		1	2	3	4	5	6	7	8	9	10	10-1 Diff	T-Stat	P
AR on Day T	Mean	-0.092%	-0.058%	-0.014%	-0.080%	0.062%	0.196%	0.198%	0.183%	0.152%	0.196%	0.288%***	14.198	0.00
	Mean (no EA)	-0.091%	-0.060%	-0.014%	-0.055%	0.065%	0.197%	0.204%	0.169%	0.149%	0.191%	0.282%***	13.945	0.00
AR on T+1 Day	Mean	-0.009%	-0.018%	-0.026%	-0.187%	-0.060%	-0.010%	0.000%	0.036%	0.013%	0.017%	0.026%*	1.354	0.09
	Mean (no EA)	-0.004%	-0.011%	0.029%	-0.142%	-0.044%	0.001%	0.008%	0.019%	0.010%	0.014%	0.054%	0.953	0.17
CAR (1,3)	Mean	-0.004%	-0.082%	-0.046%	-0.151%	-0.072%	-0.070%	-0.057%	0.039%	0.054%	0.058%	0.061%**	1.965	0.02
	Mean (no EA)	0.000%	-0.078%	-0.052%	-0.099%	-0.050%	-0.062%	-0.046%	0.023%	0.051%	0.054%	0.054%**	1.958	0.02
CAR (1,5)	Mean	-0.013%	-0.128%	-0.079%	-0.269%	-0.083%	-0.090%	-0.072%	0.042%	0.060%	0.120%	0.133%***	3.348	0.00
	Mean (no EA)	-0.009%	-0.122%	-0.083%	-0.218%	-0.066%	-0.086%	-0.064%	0.202%	0.055%	0.118%	0.127%***	3.190	0.00
No. of Observations		81,761	46,873	63,648	63,578	70,018	63,789	66,143	79,977	90,939	49,528	--	--	676, 254

This table presents the market returns of 10 decile portfolios based on StockTwits information content, where 1st decile is the stock-day observations with the most negative information, and 10th decile is the stock-day observations with the most positive information. Abnormal return and CAR are calculated using Fama & French size and valuation matched buy-and-hold abnormal return. Benchmark returns are 5(market cap) x5(book to market) portfolio returns provided on Kenneth French's personal website. Event day has been adjusted to eliminate non-trading days. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

Table 4 Firm Day StockTwits Information Content and Stock Return: Regression Analysis

	Panel A: Dependent Var: Abnormal Return on Day T+1				Panel A: Dependent Var: CAR(1,3)			
	1	2	3	4	1	2	3	4
Information Positiveness	**0.0310** ***	**0.0272** ***	**0.0226** **	**0.0313** ***	**0.0458** **	**0.0173** **	**0.0270** **	**0.0407** **
	3.02	**2.75**	**-2.21**	**2.66**	**2.08**	**2.43**	**2.19**	**2.12**
Book to Market	-0.0259 *	-0.0230	0.0470 *	0.0199 **	-0.0367	-0.0615 *	-0.0962 *	0.0296 **
	-1.95	-1.27	-1.85	2.41	-0.91	-1.80	-1.88	2.19
Size	-0.1681 ***	-0.1568 ***	-0.1369 ***	-0.0028	-0.6218 ***	-0.3573 ***	-0.3117 ***	-0.0038
	-10.02	-7.45	-4.94	-0.69	-10.8	-9.44	-5.08	-0.58
Analyst Following		-0.0763 *	-0.1271 **	-0.0102		-0.2168 **	-0.1271 **	-0.0485 **
		-1.67	-2.10	-0.87		-2.02	-2.10	-2.53
Leverage		0.0307	0.1774	-0.0175		-0.0442	0.6613	-0.0172
		0.23	1.05	-0.51		-0.16	1.58	-0.51
Profitability		-0.0104	0.1228	0.1193 *		0.2200	0.3127	0.3226 ***
		-0.08	0.64	1.87		1.49	0.62	3.12
R&D Intensity		0.0818	0.0105	0.0309		0.1535	0.1187	-0.0219
		1.09	0.04	0.39		1.66	0.94	-0.17
Sales Growth		0.0361 *	0.0003	-0.0024		0.0437 *	-0.0368	-0.0262
		1.68	0.02	-0.16		0.67	-0.75	-1.09
Dividend Pay Out		1.5967 ***	1.1701	-0.1308		0.0173 ***	0.0279	-0.0754 *
		3.37	1.56	-0.52		1.14	1.53	-1.84
SP500 Index Return			0.4158	0.4445			0.5267	0.0966
			0.54	0.74			0.83	0.98
News			-0.0276	-0.0228			-0.0743 *	-0.0596 **
			-1.03	-1.28			-1.65	-2.06
Intercept	2.4626 ***	2.5018 ***	2.5202 ***	0.0546	9.0436 ***	5.84781 ***	5.8647 ***	0.1812 **
No. of Obs	439,296	310,775	149,304	149,304	438,857	310,703	149,293	149,293
Adjusted R-Squared	1.69%	2.11%	2.15%	1.79%	1.78%	2.29%	2.23%	1.93%
Firm Fixed Effect	Yes	Yes	Yes	NO	Yes	Yes	Yes	NO
Year Fixed Effect	Yes	Yes	Yes	NO	Yes	Yes	Yes	NO

This Table presents the multivariate regression result to test the validity of StockTwits content. Abnormal return in dependent variable is the Fama-French size and market to book matched buy and hold abnormal return. Information Positiveness is the average positiveness across all users on Stocktwits for stock i on day t. For detailed textual analysis, please see section 3. All firm characteristics are winsorized at 1% level.
*,**, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

Table 5 Variance of Information and Abnormal Trading Volume

	Panel A: Dependent Var: Abnormal Trading Volume on Day T+1						Panel A: Dependent Var: Aggregated Abnormal Trading Volume of Day(1,5)					
	1		2		3		1		2		3	
Variance of StockTwits	**0.49575**	***	**0.38263**	***	**0.24863**	***	**0.82666**	***	**0.66723**	***	**0.42952**	***
Information	**26.8**		**31.27**		**23.19**		**24.46**		**27.02**		**19.10**	
Book to Market	0.01436	*	-0.00676		-0.01647	**	0.09912	***	0.00770		-0.04603	**
	1.67		-0.73		-1.95		6.31		0.41		-2.59	
Size	-0.0130	**	-0.07910	**	-0.03880	*	-0.03181	***	-0.02988	***	-0.01471	***
	-2.26		-2.27		-1.90		-3.39		-3.90		-2.59	
Analyst Following			-0.09358	***	-0.09398	***			-0.29801	***	-0.25011	**
			-3.86		-4.12				-6.08		-5.20	
Leverage			-0.28944	***	0.09283				-0.44796	***	0.42922	**
			-4.93		1.17				-3.78		2.59	
Profitability			0.46622		0.07344				-0.24592	***	0.41778	***
			1.3		1.16				-3.38		3.16	
R&D Intensity			0.01560		0.95954	***			0.21608	***	0.25560	***
			0.56		4.88				3.86		6.14	
Sales Growth			-0.02085	**	-0.03298	***			-0.00518		-0.07585	***
			-2.19		-3.05				-0.27		-3.34	
Dividend Pay Out			-0.64957		-0.74490	*			-1.73677	**	-1.10526	*
			-1.61		-1.70				-2.14		-2.29	
SP500 Index Return					-0.32746	***					-0.96870	***
					-9.96						-11.37	
News					0.31765	***					0.57047	***
					28.88						25.64	
Intercept	0.16533	***	0.46289	***	0.30998	***	0.44962	***	1.27657	***	0.73067	***
No. of Obs	337,421		273,029		111,841		449,524		315,884		150,992	
Adjusted R-Squared	2.25%		2.80%		5.13%		2.05%		4.01%		3.17%	
Firm Fixed Effect	Yes		Yes		Yes		Yes		Yes		Yes	
Year Fixed Effect	Yes		Yes		Yes		Yes		Yes		Yes	

This Table presents the regression result of abnormal trading volume and variance of StockTwits information at a firm-day level. Abnormal Trading Volume is calculated following Bamber (1996), as the difference between trading volume and the mean volume of pre-255 trading days. Divergence of StockTwits information is calculated as the standard deviation of the information of all tweets related to a stock on a given day, regardless of the sophistication levels of users.

*,**, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

Table 6 Validity Checks of Self-Reported Levels of Sophistication

Panel A Characteristics of Investor Groups: Are Self-Reported Levels Reliable?

	Novice	Intermediate	Professional	T-Test of Difference	
	Mean	Mean	Mean	3-1	
No. of Users	42,828	24,248	10,713		
	55.06%	31.17%	13.77%		
No. of Tweets	3,653,701	3,537,599	3,144,622		
	35.35%	34.23%	30.42%		
No. of Tweets Per User	85.31	145.89	293.53	208.22	***
Length of Tweets	14.72	15.14	15.29	0.56	***
No. of Professional Terms	0.715	0.763	0.806	0.091	***
Information	-0.178	-0.186	-0.161	0.017	***
Info. Precision	0.462	0.472	0.479	0.017	***
Followers	11.90	17.99	102.96	91.06	***

Table 6 Panel B Information of Different Groups and Predicting Power of Return

	Dependent Variable: Abnormal Return of (1,5) Day						
	1		2		3		
Info-Sophisticated	0.0375	***			0.0353	***	
	3.03				2.85		F-test of H_0 β(soph) = β(un-soph)
Info-Unsophisticated			0.0252	***	0.0234	***	
			2.84		2.63		
Control Variables	Yes		Yes		Yes		F-Stats
Intercept	0.0541	***	0.0547	***	0.0539		=11.64
No. of Obs	275,647		275,151		274,840		P-value
							=0.000
Adjusted R-Squared	2.09%		1.87%		2.11%		
Firm Fixed Effect	Yes		Yes		Yes		
Year Fixed Effect	Yes		Yes		Yes		

This table presents the characteristics of users with different self-reported level of sophistications to establish the construct validity of level of sophistications. Panel A provides a summary of the users' twitting activities. No. of Users is the total number of qualified users at each sophistication levels. No. Tweets is the total number of tweets posted by users of each sophistication level. Length of Tweets is the average words-per-tweet for each sophistication level. No. of Professional Terms is defined as the average number of Financial Terms used in each tweet defined by Dictionary of Financial and Business Terms (University of Toronto). Info. Precision is the standard Deviation of information within each sophistication level. Panel B reports the regression of CAR (1,5) on the information of Sophisticated and unsophisticated group. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

Table 7: Change in Information Asymmetry around MF, Univariate Results

Panel A: Change in Information Asymmetry around MF firm-day and around Non-MF firm-day

	MF-Firms	Non-MF Firms	Difference		T-Test Statistics
Change in Info Asymmetry $_{\text{Week+1}}$	0.0866	0.0000	0.0867	***	5.53
Change in Info Asymmetry $_{\text{Week+2}}$	-0.0313	0.0001	-0.0314	**	-2.08
Change in Info Asymmetry $_{\text{Week+3}}$	-0.0277	-0.0001	-0.0276	**	-1.82
Change in Info Asymmetry $_{\text{Week+4}}$	-0.0272	0.0000	-0.0272	*	-1.66

Panel B: Change in Information Asymmetry around point forecast and range forecast

	Point Forecast	Range Forecast	Difference		T-Test Statistics
Change in Info Asymmetry $_{\text{Week+1}}$	0.0831	0.1050	-0.0219	**	1.81
Change in Info Asymmetry $_{\text{Week+2}}$	-0.0590	-0.0260	-0.0330	**	-1.83
Change in Info Asymmetry $_{\text{Week+3}}$	-0.0323	-0.0269	-0.0055	**	-2.15
Change in Info Asymmetry $_{\text{Week+4}}$	-0.0553	0.0021	-0.0574		-1.31

This table presents the univariate test results of changes in information asymmetry between sophisticated and unsophisticated investors around management forecasts. Following Rogers et al. (2009), we estimate the change in information asymmetry using the following formula:

$$Change_{Week+1} = ln\left(\frac{\sum_{1}^{7} Info\ Asymmetry}{\sum_{(-7)}^{(-1)} Info\ Asymmetry}\right) \qquad Change_{Week+2} = ln\left(\frac{\sum_{8}^{14} Info\ Asymmetry}{\sum_{(-7)}^{(-1)} Info\ Asymmetry}\right)$$

$$Change_{Week+3} = ln\left(\frac{\sum_{15}^{21} Info\ Asymmetry}{\sum_{(-7)}^{(-1)} Info\ Asymmetry}\right) \qquad Change_{Week+4} = ln\left(\frac{\sum_{22}^{28} Info\ Asymmetry}{\sum_{(-7)}^{(-1)} Info\ Asymmetry}\right)$$

Panel A reports the mean change in information asymmetry around the release of Sporadic management forecasts for MF firms and Non-MF firms matched on size, Book to Market and 2-digit SIC code. Panel B reports the mean change in information asymmetry (MF firms only) around a Point Forecast and around a Range Forecast. *,**, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

Table 8: Change in Information Asymmetry around MF- Multivariate Results

Dependent Variable:	Change in Information Asymmetry - week1				Change in Information Asymmetry - week2				Change in Information Asymmetry - week3			
Variable	Coeff.		T-stat	P-Value	Coeff.		T-stat	P-Value	Coeff.		T-stat	P-Value
Management Forecast	0.0916	***	4.52	0	-0.0346	**	-2.08	0.041	-0.0335	**	-2.03	0.046
Analyst Following	-0.0122	***	-7.31	0	-0.0127	***	-5.67	0	-0.0127	***	-4.23	0
Size	0.0004		0.32	0.752	0.0004		-0.26	0.794	-0.0011		-0.38	0.704
Leverage	0.0040		0.78	0.435	0.0072		0.99	0.326	0.0105		1.02	0.312
Book to Market	0.0001		0.13	0.894	0.0013		1.57	0.121	0.0024	**	2.14	0.036
|Forecast Error|	-0.0032	**	-2.28	0.023	-0.0033	**	-2.3	0.021	-0.0031	**	-2.1	0.036
Profitability	-0.0038		-1.03	0.302	-0.0047		-1.24	0.215	-0.0051		-1.32	0.186
News	-0.0023		0.62	0.534	0.0120		0.35	0.726	0.0121		0.35	0.726
Intercept	0.0286	*	1.93	0.058	0.0408		1.59	0.116	0.0465		1.25	0.214
No. of Obs	2,519,634				2,519,634				2,519,634			
Adjusted R-Squared	3.80%				4.10%				4.00%			
Industry Fixed Effect	Yes				Yes				Yes			
Year Fixed Effect	Yes				Yes				Yes			
Error Clustered by	Industry				Industry				Industry			

This table presents the multivariate regression result of equation 1 to study the change in information asymmetry around the release of MF, controlling for firm characteristics and market characteristics. Observations include firm-day observations with and without MF releases (Management Forecast is a dummy variable set to 1 if there is MF release on the day). Dependent variables are change in information asymmetry for weeks following the release of MF compared one week prior to MF (see table 7 for detail). Industry and year fixed effect are included; errors are clustered by 2-digit SIC code.
*,**, and *** indicate significance at the 10%, 5%, and 1% level, respectively

Table 9: Change in Information Asymmetry around point and range MF- Multivariate Results

Dependent Variable:	Change in Information Asymmetry -week1				Change in Information Asymmetry - week2				Change in Information Asymmetry - week3			
Variable	Coeff.		T-stat	P-Value	Coeff.		T-stat	P-Value	Coeff.		T-stat	P-Value
Point Forecast	0.0051		0.590	0.556	-0.0101	**	-2.280	0.026	-0.0206	**	-2.390	0.017
Analyst Following	0.2971	*	1.780	0.080	0.3294	***	5.670	0.001	0.3552	***	4.110	0.005
Size	0.0593		0.700	0.488	-0.0473		-0.880	0.382	-0.0547		-0.270	0.787
Leverage	0.9140		1.650	0.104	0.5587		1.540	0.128	0.2133	*	1.667	0.094
Book to Market	0.0041		0.260	0.795	-0.0091		-0.390	0.699	-0.0179		-0.160	0.875
|Forecast Error|	0.0304		0.480	0.635	-0.0203		-0.330	0.742	0.0191		0.290	0.772
Profitability	-0.0680		-0.300	0.761	-0.0878		-0.410	0.683	-0.0728		-0.351	0.726
News	0.0123		0.550	0.537	0.0458		0.471	0.621	0.0425		0.581	0.511
Intercept	-0.8017		-1.400	0.167	-0.4017		-0.460	0.650	-0.4917		-0.350	0.726
No. of Obs	3824				3,824				3,824			
Adjusted R-Squared	1.50%				3.40%				2.47%			
Industry Fixed Effect	Yes				Yes				Yes			
Year Fixed Effect	Yes				Yes				Yes			
Error Clustered by	Industry				Industry				Industry			

This table presents the multivariate regression result of equation 2 to study the change in information asymmetry around the release of point forecast vs. range forecast, controlling for firm characteristics and market characteristics. Observations include firm-day observations when MF releases (Point Forecast a dummy variable set to 1 if a point forecast is released on the day). Dependent variables are change in information asymmetry for weeks following the release of MF compared one week prior to MF (see table 7 for detail). Industry and year fixed effect are included; errors are clustered by 2-digit SIC code. *,**, and *** indicate significance at the 10%, 5%, and 1% level, respectively

Appendix A

Variable Definitions

Variables Name	Definitions	Source
Tweet Information	Information of each tweet is calculated as natural log of number of the difference between positive words and negative words in a tweet. The classification of positive words and negative words is done with Loughran and McDonald (2011) Financial Sentiment Dictionary, Information calculation follows Antweiler and Frank (2004): $Opinion_{Tweet} = log\left(\frac{1+P}{1+N}\right)$	StockTwits
Firm-Daily Information	Firm Daily Information is the aggregated information (mean) of individual information of all tweets that mentioned stock i on day t. Information is adjusted by the trading hour: tweet posted after 4pm is adjusted to the next trading day.	StockTwits
User-Identified Information	When posting a tweet, user can voluntarily choose to disclosure her/his information on the stock mentioned, by selecting a slider between "bullish" or "bearish". In my sample, about 10% of all tweets have this information available. This information is used to check the validity of text-based information using textual analysis.	StockTwits
Reader-Identified Information	we randomly select 3,000 tweets from the sample. Risk neutral reader reads the tweets and assign the tweets into three groups: positive, neutral, negative. This information is used to check the validity of text-based information using algorithm.	StockTwits
Information Asymmetry (between sophisticated and unsophisticated investors)	Information Asymmetry is the distance (absolute value of difference) between mean information of sophisticated investors (those who identify themselves as "professional") and the mean information of unsophisticated investors (those who identify themselves as "novice") related to a given stock on a given day.	StockTwits
Changes in Information Asymmetry	Change in information asymmetry (week i) for day t, is the sum of information asymmetry from +1 day to +7 day, relative to the sum of information asymmetry from -7 day to -1 day. Following Rogers et al. (2009), formally defined as: $Change_{Week1} = ln\left(\frac{\sum_{1}^{7} Info\ Asymmetry}{\sum_{(-7)}^{(-1)} Info\ Asymmetry}\right)$	StockTwits

StockTwits Information Divergence	StockTwits Information Divergence is the standard deviation of all users' information related to a given stock on a given day, regardless of the sophistication level of users.	StockTwits
Information Precision of Investor Group	Information precision of investor group is the standard deviation of information related to a given stock on a given day given by users of one level of sophistication.	StockTwits
Point Forecast	Point Forecast is a dummy variable that equals to 1 when the second estimation value in the first call detail tape is missing, indicating that manager only provides on estimation for future earnings. This classification is double checked by examining the textual presentation in the "comment" section.	First Call
Sporadic Forecast	Sporadic Forecast is identified by first eliminating the MF made within -3 to +3 days of an earnings announcements (identified using Compustat and IBES, which ever date is earlier). Second, we look at two consecutive years and in which week the sample forecast happened, if it happens in the same week of year t and year t-1, then both forecasts will be considered "regular" and dropped.	First Call
Raw Return	Raw return is the RET from CRSP data base	CRSP
Abnormal Return of Day t	Abnormal Return is the raw return of firm i on day t and the benchmark return of firm i's matched portfolio based on (5x5 size and market-to-book ratio). Benchmark portfolio is formed following Fama French method, benchmark return is acquired from Kenneth French's personal website.	CRSP/ Kenneth French
CAR[i,j]	CAR(i,j) is defined as the difference between the buy and hold return (BAHR) for stock k from day i to day j, and the buy and hold return of the benchmark group for the same period.	CRSP
Abnormal Trading Volume	Abnormal Trading Volume is the natural log of trading volume on day t minus its average trading volume for the previous 255 trading days.	CRSP
Trading Volume	Trading Volume is the VOL from CRSP data base	CRSP
Absolute Value of S&P 500 Daily Change	Absolute Value of S&P 500 Index Daily Change	CRSP
Book to Market	BOOKTOMARKET is the book value of equity divided by the book value of equity. Ratio of book value of common equity to market capitalization (CEQQ/[PRCCQ *	Compustat/ CRSP

	CSHOQ]), If CEQQ is missing, book value of common equity will be calculated as ATQ-LTQ	
Size	The natural log of lagged market capitalization on the day of last annual report	Compustat/ CRSP
Leverage	LEVERAGE is the ratio of long term liability to total assets from most recent balance sheet disclosures.	Compustat
Dividend Payout	Dummy Variable if firm t has dividend paid to the investors within the last fiscal year	Compustat
R&D Intensity	R&D intensity is an indicator variable if the ratio of R&D expense to total expense falls in the upper quartile of sample firms.	Compustat
Analyst Following	ANALYST FOLLOWING is the natural log of 1+ latest number of analysts following the firm.	I/B/E/S
Profitability	Profitability is measured as EBITDA of a firm in the last annual report, divided by the total value of assets in the last annual report.	Compustat
Sales Growth	Sales Growth is the year-to-year change in the total revenue reported in latest annual report	Compustat
Forecast Error	Mean of the Absolute value of the forecast error (actual EPS-consensus EPS) for the past 12 fiscal quarters, scaled by stock price.	I/B/E/S
News	Natural log of number of news articles on Factiva	Factiva
Industry Dummy	2-digit SIC classification	Compustat

Bibliography

Ali, Ashiq, Lee-Seok Hwang, and Mark A. Trombley. 2000. Accruals and future stock returns: Tests of the naïve investor hypothesis. *Journal of Accounting, Auditing & Finance* 15, (2): 161-81.

Antweiler, Werner, and Murray Z. Frank. 2004. Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*59, (3): 1259-94.

Atiase, Rowland K., and Linda Smith Bamber. 1994. Trading volume reactions to annual accounting earnings announcements: The incremental role of predisclosure information asymmetry. *Journal of Accounting and Economics* 17, (3): 309-29.

Ayers, Benjamin C., Oliver Zhen Li, and P. Eric Yeung. 2011. Investor trading and the post-earnings-announcement drift. *The Accounting Review* 86, (2): 385-416.

Baginski, Stephen P., Edward J. Conrad, and John M. Hassell. 1993. The effects of management forecast precision on equity pricing and on the assessment of earnings uncertainty. *Accounting Review*: 913-27.

Bamber, Linda Smith, Orie E. Barron, and Douglas E. Stevens. 2011. Trading volume around earnings announcements and other financial reports: Theory, research design, empirical evidence, and directions for future research. *Contemporary Accounting Research* 28, (2): 431-71.

Bamber, Linda Smith, Orie E. Barron, and Thomas L. Stober. 1999. Differential interpretations and trading volume. *Journal of Financial and Quantitative Analysis* 34, (03): 369-86.

Barron, Orie E., Donal Byard, and Oliver Kim. 2002. Changes in analysts' information around earnings announcements. *The Accounting Review* 77, (4): 821-46.

Bartov, Eli, Suresh Radhakrishnan, and Itzhak Krinsky. 2000. Investor sophistication and patterns in stock returns after earnings announcements. *The Accounting Review* 75, (1): 43-63.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2, (1): 1-8.

Collins, Daniel W., Guojin Gong, and Paul Hribar. 2003. Investor sophistication and the mispricing of accruals. *Review of Accounting Studies* 8, (2-3): 251-76.

Coller, Maribeth and Yohn Lombardi, Journal of Accounting Research Vol. 35, No. 2 (autumn, 1997), pp. 181-191

Cready, William, Abdullah Kumas, and Musa Subasi. 2014. Are trade Size-Based inferences about traders reliable? evidence from institutional Earnings-Related trading. *Journal of Accounting Research* 52, (4): 877-909.

Chen, Hailiang, Prabuddha De, Yu Jeffrey Hu, and Byoung-Hyoun Hwang. 2014. Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies* 27, (5): 1367-403.

DellaVigna, Stefano, and Joshua M. Pollet. 2009. Investor inattention and friday earnings announcements. *The Journal of Finance* 64, (2): 709-49.

Diamond, Douglas W., and Robert E. Verrecchia. 1991. Disclosure, liquidity, and the cost of capital. *The Journal of Finance* 46, (4): 1325-59.

Frazzini, Andrea, Ronen Israel, and Tobias J. Moskowitz. 2012. Trading costs of asset pricing anomalies. *Fama-Miller Working Paper*: 14-05.

Giannini, Robert Charles, Paul J. Irvine, and Tao Shu. 2014. Do local investors know more? A direct examination of individual investors' information set. Paper presented at A Direct Examination of Individual Investors' Information Set (August 8, 2014). Asian Finance Association (AsFA) 2013 Conference, .

Grossman, Sanford J., and Joseph E. Stiglitz. 1980. On the impossibility of informationally efficient markets. *The American Economic Review*: 393-408.

Hasbrouck, Joel. 1991. The summary informativeness of stock trades: An econometric analysis. *Review of Financial Studies* 4, (3): 571-95.

Hirshleifer, David A., James N. Myers, Linda A. Myers, and Siew Hong Teoh. 2008. Do individual investors cause post-earnings announcement drift? direct evidence from personal trades. *The Accounting Review* 83, (6): 1521-50.

Hirst, D. Eric, Lisa Koonce, and Shankar Venkataraman. 2008. Management earnings forecasts: A review and framework. *Accounting Horizons* 22, (3): 315-38.

Holthausen, Robert W., and Robert E. Verrecchia. 1990. The effect of informedness and consensus on price and volume behavior. *Accounting Review*: 191-208.

Indjejikian, Raffi J. 1991. The impact of costly information interpretation on firm disclosure decisions. *Journal of Accounting Research*: 277-301.

Kim, Oliver, and Robert E. Verrecchia. 1997. Pre-announcement and event-period private information. *Journal of Accounting and Economics* 24, (3): 395-419.

Kim, Oliver, and Robert E. Verrecchia. 1994. Market liquidity and volume around earnings announcements. *Journal of Accounting and Economics* 17, (1): 41-67.

Kim, Oliver, and Robert E. Verrecchia. 1991. Market reaction to anticipated announcements. *Journal of Financial Economics* 30, (2): 273-309.

Lambert, Richard, Christian Leuz, and Robert E. Verrecchia. 2007. Accounting information, disclosure, and the cost of capital. *Journal of Accounting Research* 45, (2): 385-420.

Lee, Charles MC, Belinda Mucklow, and Mark J. Ready. 1993. Spreads, depths, and the impact of earnings information: An intraday analysis. *Review of Financial Studies* 6, (2): 345-74.

Leuz, Christian, and Robert E. Verrecchia. 2000. The economic consequences of increased disclosure (digest summary). *Journal of Accounting Research* 38, : 91-124No.

Lev, Baruch. 1988. Toward a theory of equitable and efficient accounting policy. *Accounting Review*: 1-22.

Loughran, Tim, and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-Ks. *The Journal of Finance* 66, (1): 35-65.

Miller, Brian P. 2010. The effects of reporting complexity on small and large investor trading. *The Accounting Review* 85, (6): 2107-43.

Miller, Edward M. 1977. Risk, uncertainty, and divergence of opinion. *The Journal of Finance* 32, (4): 1151-68.

Pownall, Grace, Charles Wasley, and Gregory Waymire. 1993. The stock price effects of alternative types of management earnings forecasts. *Accounting Review*: 896-912.

Rogers, Jonathan L., Douglas J. Skinner, and Andrew Van Buskirk. 2009. Earnings guidance and market uncertainty. *Journal of Accounting and Economics*48, (1): 90-109.

Tang, Michael Minye. 2014. Consistency in management earnings guidance patterns. *Available at SSRN 1952804*.

Utama, Siddharta, and William M. Cready. 1997. Institutional ownership, differential predisclosure precision and trading volume at announcement dates.*Journal of Accounting and Economics* 24, (2): 129-50.

Walther, Beverly R. 1997. Investor sophistication and market earnings expectations. *Journal of Accounting Research*: 157-79.

Watts, Ross L., and Jerold L. Zimmerman. 1979. The demand for and supply of accounting theories: The market for excuses. *Accounting Review*: 273-305.

Yohn, Teri Lombardi. 1998. Information asymmetry around earnings announcements. *Review of Quantitative Finance and Accounting* 11, (2): 165-82.