Local Twitter Activity and Stock Returns*

Bok Baik
Seoul National University
Email: bbaik@snu.ac.kr

Qing Cao
Texas Tech University
Email: qing.cao@ttu.edu

Sunhwa Choi
Lancaster University
Email: s.choi@lancaster.ac.uk

Jin-Mo Kim
Rutgers University
Email: kimjm@business.rutgers.edu

July 2015

* We thank Brian Bushee (FARS discussant), Linda Myers, Gaizka Ormazabal Sanchez, Steven Chong Xiao, Steven Young, and seminar participants at Korea University, Lancaster University, Manchester University, Yonsei University, Seoul National University, SKK GSB, and the 2015 Financial Accounting and Reporting Section (FARS) Midyear Meeting for their helpful comments and suggestions.

Local Twitter Activity and Stock Returns

ABSTRACT

Using geographic proximity as a measure of information advantages of social media users, this paper examines the informational role of social media in stock markets. We find that the tone in tweets by local Twitter users predicts future stock returns; in contrast, such predictability does not exist for nonlocal users. Moreover, the positive relation between the tone in local tweets and stock performance is more salient in firms with high information asymmetry. We also find that the tone of local tweeter users predicts subsequent earnings announcement returns. These findings suggest that geographic proximity indicates the availability of private information and social media can provide value-relevant information about local companies.

Keywords: Twitter; Social media; Textual analysis, Geographic proximity; Stock return.

1. Introduction

The importance of social media in financial markets has increased substantially over the past decade. Firms use social media outlets such as Twitter and Facebook to disseminate corporate information (SEC, 2013; Blankespoor, Miller, and White, 2014). Large institutional investors such as hedge funds use the social media data to develop trading strategies. For example, the world's largest hedge fund, Bridgewater, uses social media to model economic activity in real-time.[1] However, few studies examine social media's return predictability and it is unclear whether the activities on social media have any investment value. For example, Bollen, Mao, and Zeng (2011) show that the public sentiment of Twitter users predicts Dow Jones Industrial Average (DJIA) returns.[2] Chen, De, Hu, and Hwang (2014) find that the negative view in articles on Seeking Alpha, an investment-related social media platform, predicts future stock returns. However, most studies (e.g., Antweiler and Frank 2004; Das and Chen 2007; and Sprenger, Tumasjan, Sandner, and Welpe 2013) suggest that social media activities are not significantly related to future returns.

To explore the informational role of social media, we focus on the role of geographic proximity in social media activities. There is a growing body of literature on the geography of investment and geographically proximate investors' information advantages. For example, Coval and Moskowitz (1999) find that U.S. fund managers exhibit a strong preference for locally headquartered firms. Coval and Moskowitz (2001) show that U.S. fund managers earn an additional annual return of 2.65% from their local investments compared to their nonlocal investments, providing evidence of local advantages. More importantly, there could be

[1] http://www.zerohedge.com/news/2013-12-12/worlds-largest-hedge-fund-uses-twitter-real-time

[2] This research led to some hedge funds adopting a trading strategy to track the public mood in Twitter to predict market movements (Financial Times, 2011).

significant interactions between social media and geographic proximity. On the one hand, it is unclear ex ante whether physical distance matters in social media activities, where communications around the world can be real-time and it is not uncommon to build social networks with someone from other countries. On the other hand, Takhteyev, Gruzd, and Wellman (2012) show that a substantial share of Twitter networks connect users within the same regional cluster, typically the size of a metropolitan area, suggesting that geographical proximity also affects social interactions via social media networks. To the extent that social media users exhibit local preference in their social media activities, we expect geographic proximity to play an important role in social media activities, thus providing an ideal setting for an examination of the informational role of social media in financial markets.

In this study, we use Twitter, an online social networking and microblogging platform, as the main outlet of social media activities. Twitter has been growing in popularity as a venue for communication among individual investors. Twitter has been at the top of the micro blogging platforms since 2006, and as of the first quarter of 2014, Twitter had 255 million monthly active users who posted 500 million tweets each day. Twitter users can post short messages called 'tweets' with up to 140 characters. These messages are usually shared publicly or within a network of followers. Many of the messages are discussions of stocks and trading (Sprenger, Tumasjan, Sander, and Welpe, 2013).

Compared to other social media platforms, Twitter has several unique features that facilitate the quick diffusion of information and interactive conversations among users, thus making it ideal for examining the role of social media activities in capital markets. For example, Twitter's word limit of up to 140 characters makes messages short and frequent (rather than lengthy and infrequent as in in blogs or internet message boards), quickly grabbing people's

attention. Twitter shows others' new messages in the main pages when a Twitter user follows other users, enabling users to be continuously informed of others' updates. Twitter users can 'retweet' messages by reposting others' tweets. Twitter users can also use the hashtag symbol (#) before a relevant keyword to categorize tweets and improve their accessibility for Twitter searched. Unlike other social networking services such as Facebook, Twitter requires no reciprocal relationship. In other words, a user can follow any other user and the user being followed need not follow back, so networks in Twitter are driven by information rather than based on personal relationships.

Our sample consists of 552,012 messages on Twitter from July 2011 through March 2012 about 646 firms. [3] Employing the location based (geo-code) search query from the Twitter application programming interface (API), we identify local Twitter activity. We define local (nonlocal) users as those located within the same (different) state as the firm's headquarters (Baik, Kang, and Kim, 2010). We measure the negative (positive) tone of the Twitter messages as the fractions of negative (positive) words to total words (Loughran and McDonald, 2011).

We find that local tweets account for an average of 65% of total tweets, suggesting a significant local bias in social media activities. We also find that local Twitter users tend to tweet about firms with high information asymmetry, including firms with low book-to-market, less liquid firms, non-S&P 500 firms, R&D intensive firms, firms with low analyst coverage.

Next, we find that the views expressed in Twitter messages predict future stock returns. Specifically, we find that the fraction of negative words about a firm in tweets predicts future stock returns. More importantly, when we partition local and nonlocal tweets based on the geographic proximity of Twitter users to the company, we find that the fraction of negative

[3] We eliminate Twitter accounts managed by firms themselves.

words in local tweets predicts future returns, while such relation does not exist for nonlocal tweets. Specifically, the next-day (the next ten-day) abnormal return is 0.04% (0.14%) lower when the fraction of negative words in local tweets is one standard deviation higher. In addition, the relation between the negative tone in local tweets and future returns is not reversed in the subsequent periods. To the extent that Twitter users located near firms have better access to information about local firms than do remote users, these results highlight the ability of local social media activities to forecast future stock returns.

We also find that the ability of local tweets to predict returns comes primarily from firms with high information asymmetry. Specifically, the relation between the tone in local tweets and stock returns is more evident for small firms, firms with low analyst following, and young firms. This result is consistent with previous studies that have documented the information advantage of local investors (Coval and Moskowitz, 2001; Malloy, 2005; Ivkovic and Weisbenner, 2005; Baik, Kang, and Kim, 2010).

Finally, consistent with information advantages of local social media users, we find that the fraction of negative words in local tweets predicts future earnings announcement returns while that in nonlocal tweets does not.

Overall, these results suggest that local Twitter users have information advantages about local stocks and the views they express in tweets about firms' prospects has return predictability.

We extend the existing literature in three important ways. First, our paper sheds new light on the literature on the investment value of social media activities (Bollen, Mao, and Zeng, 2011; Giannini, Irvine, and Shu, 2013; Chen, De, Hu, and Hwang, 2014). We provide evidence that the aggregated view from social media has investment value and this investment value stems primarily from local users' superior knowledge. Our findings thus suggest that local social media

cn be used to predict returns. Our study is closely related to the concurrent work by Gianni, Irvine, and Shu (2013). While they show that messages on StockTwits.com by local investors do not contain value-relevant information, we present evidence that local Twitter activities predict future stock returns. [4]

Second, recent empirical evidence suggests that geographically proximate investors have significant information advantages over distant investors (Coval and Moskowitz, 1999, 2001; Ivkovic and Weisbenner, 2005). Our paper shows that the tone in tweets by local Twitter users predicts future stock returns and such return predictability is more pronounced in stocks with high information asymmetries, suggesting that local social media users communicate information that is not yet impounded into stock prices. Our analysis offers additional evidence on the role of geography in capital markets.

Finally, we also add to the literature on textual analysis of media (Das and Chen, 2007; Tetlock, 2007; Tetlock, Saar-Tsechansky, and Macskassy, 2008). We extend the literature by showing that textual analysis can be used to extract value-relevant information from a large set of individual messages posted on social media.

The rest of this paper proceeds as follows. In Section 2, we review the related literatures on local advantages and the use of qualitative information from media in investments. Section 3 describes the data and summary statistics. In Section 4, we provide empirical evidence on the relation between the tone of local and nonlocal tweets and future stock returns. Section 5 presents the results from robustness tests. Finally, we present summary and concluding remarks in Section 6.

[4] Please see Section 2 for further discussion.

2. Literature Review

Geographic Proximity and Information Advantages

Prior research shows that geographic proximity affects the amount of information available to investors. For example, Coval and Moskowitz (1999) analyze the role of geographic proximity in the context of U.S. mutual fund managers and show that U.S. fund managers exhibit a bias toward locally headquartered firms, particularly with small, highly leveraged firms that produce nontraded goods. Coval and Moskowitz (2001) also show that on average fund managers generate an additional annual return of 2.65% from their local investments compared to their nonlocal investments. They argue that fund managers earn such abnormal returns from their local holdings because they acquire better information about local companies. Ivkovic and Weisbenner (2005) find that the average U.S. household generates an additional annual return of 3.2% from its local holdings relative to its nonlocal holdings.[5] Similarly, Baik, Kang, and Kim (2010) provide evidence that geographically proximate institutional investors execute more profitable trades.

Several papers also investigate the relation between distance and analyst performance. Malloy (2005) shows that geographically proximate analysts in the U.S. issue more accurate earnings forecasts, update their forecasts more frequently, and have a greater impact on stock prices, suggesting that geographically proximate analysts possess an information advantage over other analysts. Bae, Stulz, and Tan (2008) also find evidence of local analysts' information advantages for a sample of 32 non-U.S. countries.

[5] However, Seasholes and Zhu (2010) show that individual investors' portfolios of local holdings do not generate abnormal performance and their purchases of local stocks significantly underperform sales of local stocks. Thus, they conclude that individual investors do not have value-relevant information about local stocks.

These findings suggest that geographically proximate social media users may have significant information advantages over remote users. Local social media users can have better access to information about local firms than nonlocal users can. Local social media users can follow the firm through local media reports and they may have access to other sources of information about local companies, such as employees, managers, suppliers, and customers. As Ivkovic and Weisbenner (2007) suggest, information about local firms can be quickly diffused among local individuals through social interactions in their local areas or through social media activities.

Qualitative Information from Media

There is a growing literature on the relation between qualitative information in media and financial market variables. Tetlock (2007) links the *Wall Street Journal*'s popular "Abreast of the Market" column with subsequent stock returns and trading volume and finds that high levels of pessimistic words predicts lower returns the next day. Tetlock, Saar-Tsechansky, and Macskassy (2008) find that the fraction of negative words in news stories predicts subsequent earnings and stock returns, suggesting that linguistic media contents capture firms' fundamentals.

Besides traditional media, recent studies examine the role of social media outlets in the capital market.[6] In an early study, Wysocki (1998) examines the message volume of 50 firms with the highest activity on Yahoo! message boards and finds that changes in overnight message posting volume predict changes in next-day trading volume and abnormal returns. Das and Chen (2007) develop computational linguistics techniques to extract investor sentiment from stock

[6] There are a few studies on firms' use of social media. For example, Blankespoor, Miller, and White (2014) provide evidence that tweets facilitate the dissemination of news, particularly for firms with high information asymmetry. Chen, Hwang, and Liu (2013) show that CEOs/CFOs' tweets contain value-relevant information. They document that the tone in tweets by an executive working for a publicly traded company predicts the firm's future stock performance. In this study, we focus on the social media activities by individual users, rather than by firms.

message boards and apply the algorithm for 24 stocks in high-tech industry. They find that the sentiment weakly predicts future stock returns at the aggregate levels but there is no relation between sentiment and stock prices at the individual stock level. Antweiler and Frank (2004) also examine messages posted on Yahoo! Finance and Raging Bull about 45 firms and measure the bullishness of the messages contents based on computer-based algorithm. They find that bullishness is significantly related to contemporaneous returns but not to subsequent returns. Chen, De, Hu, and Hwang (2014) examine whether the view expressed in Seeking Alpha, an investment-related social media platform, predicts future returns. They find that the fraction of negative words in articles and the fraction of negative words in comments written in response to the articles both negatively predict future stock returns and earnings surprises.[7]

Bollen, Mao, and Zeng (2011) is the first study to use Twitter as a predictor of stock market performance. Measuring public sentiment from a collection of public tweets, they find that a "calm" mood of Twitter users predicts the Dow Jones Industrial Average (DJIA) returns. At the individual stock level, Sprenger, Tumasjan, Sandner, and Welpe (2013) analyze 250,000 tweets for firms included in the S&P 100 index on a daily basis from January 2010 to June 2010. Their findings show a significant relation between sentiment and contemporaneous returns but no relation between sentiment and future returns. Giannini, Irvine, and Shu (2013) use a dataset from StockTwits.com, a social media platform to post messages about stocks. They use 216,266 messages posted on StockTwits.com from July 2009 to June 2011 and collect users' location information from their profile pages. Their results show that users' evaluation is *negatively* associated with future returns, suggesting underperformance of StockTwits users. However, this

[7] It is not obvious why people share their value-relevant information with others through various social media activities. Users might obtain significant utility from the attention and recognition from their posting. It is possible that users can get feedback from interaction with other users (Chen, De, Hu, and Hwang, 2014).

underperformance is not present for local users, who are located within 100 miles of the corporate headquarters. Our study is different from Giannini, Irvine, and Shu (2013) in several ways. First, the users of StockTwits.com consist of only a small subset of Twitter users who are particularly interested in stock investing.[8] On the other hand, our data, directly from Twitter, reflect more general user groups of social media.[9] Second, as their location information is extracted from users' public profile pages, they could get this information for only one third of their original sample. In contrast, we identify the location of Twitter users employing location (geo-code)-based search query from Twitter Search API regardless of whether they make their locations publicly available on the profile page, thereby retaining all available samples. Third, their empirical results show that the difference between local and nonlocal users is entirely driven by nonlocal users' underperformance rather than by local users' superior stock-picking ability. In other words, the tone in tweets sent by local users does not have incremental information relative to the average user. By contrast, our results indicate that the fraction of negative words in local tweets negatively predicts future returns, consistent with the information advantage view of local users.

3. Data

Data

[8] Wang et al. (2014) report that the number of active users (i.e., who have posted at least one message during the period 2009-2014) and registered users at StockTwits are 86,497 and 300,000, respectively.

[9] StockTwits was originally built on Twitter when founded in 2008 by introducing the use of a cashtag ($TICKER) to track stock symbols. In 2009, it established a separate platform (and separate company), in which users can link their StockTwits accounts to their Twitter and vice versa. In June 2012, Twitter also introduced this cashtag feature and blocked posting to StockTwits accounts via Twitter from March 2013, weakening the link between the two services. The co-founder of StockTwits described this as Twitter hijacking their services (CNN, 2012).

We initially choose six industries including Pharmaceutical Preparation Manufacturing, Retail Trade, Software Publishers, Savings Institutions, Health Care and Social Assistance, and Travel, Accommodation & Food Services, in which social media activities are active (Constantinides, Romero, and Boria, 2009; Culnan, Patrick, McHugh, and Zubillaga, 2010; Greene and Kesselheim, 2010; Noone, McGuire, and Rohlfs, 2011; Phil and Nguyen, 2013; Yu, Duan, and Cao, 2013). We then randomly select 1,044 firms from the industries using stratified sample method and collect the daily Twitter activities for these firms for the 9-month period from July 1, 2011 to March 31, 2012. A web crawler was developed to download data including Twitter username, user id, location (user self-entered), the date-time of tweets, submission type (tweet or re-tweet), the text contents for each tweet, and source used by the user to tweet. To avoid spam messages and other advertising tweets, we filter out tweets that include URLs only. (Please see appendix I for details of data collection.)

We then classify tweets into local or nonlocal tweets based on users' location. Specifically, a tweet is classified as a local (nonlocal) tweet if the user is located in the same (different) state(s) as the firm's headquarters (Baik, Kang, and Kim, 2010). We also collect the data on the conventional news media such as newspapers and business magazines via the Google News, which aggregates comprehensive news articles from numerous sources. We select 10 news sources including ABC News, New York Times, Reuters, USA Today, Fox News, Wall Street Journal, Washington Post, CNN, Economist, and Forbes as the conventional news media outlets. We also classify a newspaper or magazine article as local or nonlocal in a similar way.[10]

[10] We acknowledge that our classification of local versus nonlocal news does not really reflect the information advantage of local newspapers because these 10 outlets are all national press rather than daily newspaper of a city.

To construct the tone of tweets, we follow prior literature and use the frequency of negative and positive words (sentiment lexicons) in the tweet messages. We use the word list compiled by Loughran and McDonald (2011), which is specifically designed for the financial market studies. Loughran and McDonald (2011) show that word lists developed for other disciplines likely misclassify common words in financial text. They find that three-fourths of the words identified as negative by the widely used Harvard Dictionary are words typically not considered negative in financial contexts.

For each firm, we count the number of positive and negative words across all the tweets on the day and then divide them by the number of total words in the tweets to create two word classifications, positive and negative words (*PosTwt* and *NegTwt*). Thus, our variables of the tone in tweets aggregate all the Twitter activities about the firm during the day. Since prior studies suggest that bad news contains more value-relevant information (e.g., Tetlock, 2007; Tetlock, Saar-Tsechansky, and Macskassy, 2008), we expect that the negative tone is more likely to predict stock returns than the positive tone. We measure the tone of local (nonlocal) tweets as positive or negative words in all the local (nonlocal) tweets about the firm for each day (*Local* (*Nonlocal*) *PosTwt* and *Local* (*Nonlocal*) *NegTwt*). For illustration, Appendix II provides examples of local and nonlocal tweets about the same event but with different tone. The tone in the conventional news media is similarly measured.

We then merge the Twitter activity data for 1,044 firms over the July 2011-March 2012 period (287,085 firm-day observations) with Compustat and CRSP. We exclude non-U.S. firms and observations with missing variables for our analyses. We also require that a firm has at least one Twitter message on a given day to ensure that the tone variables can be defined. As a result, as described Table 1, our final sample for the return prediction test consists of 63,103 firm-day

observations (646 unique firms). This sample includes 552,012 individual Twitter messages with 357,592 (194,420) messages classified as local (nonlocal), implying that about 65% of our sample are local tweets. This high ratio shows Twitter users' strong tendency to cover local firms.

Descriptive Statistics

Table 2 provides descriptive statistics for the variables used for our analysis. The mean of the number of tweets for the day (*CountTwt*) is 8.75, indicating that the average sample firm is discussed 8.75 times per day in Twitter. Out of 8.75 messages, 5.67 messages are created by local users (*Local CountTwt*) and 3.08 messages are tweeted by nonlocal users (*Nonlocal CountTwt*). The ratio of the number of positive and negative words to the total number of words in tweet messages is 0.2% and this ratio is not statistically different between local and nonlocal tweets. In comparison, a firm is typically covered by the conventional news media 0.62 times a day, with 0.44 (0.18) times appearance in local (nonlocal) papers. The average fraction of negative (positive) words in the conventional news media is 0.2% (0.6%). Note that the numbers of observations with available news-related variables are substantially smaller than those with the Twitter variables. For example, there are 10,299 firm-day observations with available nonlocal tone variables in news, which is only 16% of our sample. Thus, in the subsequent analyses, we set missing values for the news-related tone variables as zero unless otherwise indicated, to avoid the loss of sample.

Table 3 presents correlations among the key variables. The correlation between positive tone (*PosTwt*) and negative tone (*NegTwt*) in the tweet messages is relatively low (0.07). The correlation between the positive (negative) tone in local tweets and the positive (negative) tone in nonlocal tweets is 0.55 (0.52). While the fraction of positive words in tweets (*PosTwt*) is not

significantly related to abnormal returns on the following day (*AR(t+1)*), the negative tone in tweets (*NegTwt*) is negatively related to the next-day returns (*AR(t+1)*) at the 5% level. These correlation results imply that the frequency of negative words captures the tone of the text (Tetlock, 2007; Tetlock, Saar-Tsechansky, and Macskassy, 2008; Loughran and McDonald, 2011; Kim and Meschke, 2011). As expected, when the tweets are split into local and nonlocal, this negative relation between the negative tone and next-day abnormal returns is observed only for local tweets (*Local NegTwt*) but not for nonlocal tweets (*Nonlocal NegTwt*). This result indicates that the tone in local tweets has return predictability in general.

4. Main Empirical Results

Determinants of Twitter Coverage and Local Twitter Coverage

To examine the cross-sectional determinants of Twitter coverage, we estimate cross-sectional regressions of Twitter coverage on firm characteristics. Following previous studies (Wysocki, 1998; Baik, Kang, and Kim, 2010; Bushee, Core, Guay, and Hamm, 2010), we include firm size, market-to-book ratio, return volatility, share turnover, stock price, S&P 500 inclusion, abnormal returns during the recent periods, firm age, dividend yield, the ratio of research and development (R&D) expenses, and the number of analyst following as potential determinants of Twitter coverage. Note that we use a larger sample here for the Twitter coverage determinant analysis because firms that do not have Twitter coverage on a day (i.e., zero coverage) are also included in the analyses.

In Column (1) of Table 4, we estimate a logit regression of an indicator variable for daily Twitter coverage on these determinant variables. The regression result shows that Twitter users are more likely to send a tweet message about large firms, value firms (i.e., firms with high

book-to-market ratios), firms with high return volatility and share turnover, firms that are included in the S&P 500 index, firms that have experienced positive returns during the recent periods, mature firms, firms with high dividend yield, firms with lower R&D activities, and firms with greater analyst following. We next replace the indicator variable with the number of Twitter messages (i.e., count) and estimate the negative binomial regression to account for the count-data nature of the dependent variable (Rock, Sedo, and Willenborg, 2000). The result in Column (2) shows very similar finding to that from the logit regression in Column (1).

To test whether the determinants for local tweets differ from those for nonlocal tweets, we then examine the determinants of the Twitter coverage by local users compared to nonlocal users for those with non-zero Twitter coverage. For the sample that has non-zero Twitter coverage, we define an indicator variable that takes the value of 1 if the firm is covered by local Twitter users but not by nonlocal Twitter users, and 0 otherwise (i.e., covered by both local and nonlocal Twitter users or covered only by nonlocal Twitter users). We then estimate a logit regression to examine firm characteristics associated with cases in which only local Twitter users send a tweet message about a firm.

The result in Column (3) shows that local Twitter users are more likely to tweet about firms with low book-to-market ratio (i.e., growth firms), firms with low turnover, firms with high stock prices, firms that are not included in the S&P 500 index, firms that have had low stock returns during the recent periods, firms with low dividend yield, firms with high R&D activities, and firms with low analyst following. These firm characteristics are generally associated with high information asymmetry. To the extent that local Twitter users have an information advantage over nonlocal Twitter users, the results suggest that Twitter users are more likely to

tweet about local companies with high information asymmetry in which they are more likely to be informed.

In Column (4), we use the ratio of local tweet counts to total tweet counts as the dependent variable and estimate a Tobit regression since the value of the dependent variable is bounded between 0 and 1. While the results are similar to those in Column (3), there are two differences. The coefficient on *Age* is now significantly negative, consistent with local Twitter users selecting firms with high information asymmetry (i.e., younger firms), while the coefficient on *R&D* is now insignificant in this specification.

Overall, the results in Table 4 suggest that local Twitter users are more likely to tweet about firms that have high information asymmetry in which they have a relative information advantage.

Predicting Subsequent Stock Returns with the Tone in Tweets

In this section, we examine whether the tone in Twitter messages predicts future returns. To the extent that Twitter messages convey new information about firms' prospects, we expect that the tone in tweets, particularly negative news, will be related to future stock returns. Based on the literature on geographic proximity and information advantages, we also predict that predictability of the tone in tweets, if any, will be more pronounced in tweets made by local users.

We measure a firm's abnormal returns on day t+1, where day t is the day of tweets or the ensuing trading day if the tweet is on a non-trading day. Abnormal returns are calculated as raw returns adjusted for 25 (5*5) value-weighted portfolios by size and book-to-market ratio (Fama and French, 1996). The positive and negative words in local and nonlocal tweets are our variables of interest. To control for the tone in the conventional news media, we include positive

and negative word categories in the regression. Following previous studies (Tetlock, 2007; Tetlock, Saar-Tsechansky, and Macskassy, 2008; Chen, De, Hu, and Hwang, 2014), we include several control variables to ensure that the predicting ability of the tone in tweets is incremental to other firm characteristics that might affect stock returns. We include four control variables for a firm's recent returns, the cumulative abnormal return from the (-30, -3) trading day window and the abnormal returns on day -2, -1, and 0. These variables are expected to capture the return predictability of past returns (Jegadeesh and Titman, 1993). In particular, including the abnormal return on the day of Twitter message (day 0) effectively controls for a possibility that Twitter messages merely reflect the news that are released in the market on the same day. In addition, we control for the firm's most recent earnings surprises, firm size, book-to-market ratios, and share turnover (Tetlock, Saar-Tsechansky, and Macskassy, 2008; Chen, De, Hu, and Hwang, 2014).

In Panel A of Table 5, we report the regression results with the next-day abnormal returns as the dependent variable. In Column (1), we use positive and negative word categories in all tweets (*PosTwt* and *NegTwt*) as the key independent variables to examine the predictability of the overall tone in tweets about the firm. The coefficient on *NegTwt* is significantly negative at the 10% level, suggesting that the negative news about the firm expressed in the tweet messages predicts the next-day stock returns. On the other hand, the coefficient on *PosTwt* is not significant, consistent with prior studies that the relation between positive words and stock returns is weak (e.g., Tetlock, 2007; Tetlock, Saar-Tsechansky, and Macskassy, 2008).

In Column (2), we use positive and negative word categories in local tweets as the independent variables and find that the negative news in local tweets (*Local NegTwt*) is associated with lower stock return on the following day. Again, the positive tone is not significantly related to stock returns. In contrast, when positive and negative words in nonlocal

tweets are used in Column (3), they are all insignificant. We include the tone of local tweets and nonlocal tweets together in Column (4) and confirm that the ability of the negative news in tweets to predict the next-day returns comes entirely from local Twitter users, consistent with information advantages of local social media activities. The magnitude of the coefficient suggests that the next-day abnormal returns are 0.04% lower when the fraction of negative words in local tweets is one standard deviation higher.

Turning to the control variables, the tone in the conventional news media is not significantly related to future stock returns for our sample.[11] The coefficients on *AR(t+0), AR(t-1), AR(t-2)* are all significantly negative, consistent with a short-term reversal.

In Panel B of Table 5, we repeat the analyses after replacing the next-day return with the next three-day returns (i.e., cumulative abnormal returns over the (t+1, t+3) window). The results are similar to those in Panel A. One exception is that the fraction of negative words in nonlocal tweets is negatively related to the next three-day returns in Column (3) in which the tone variables of local tweets are not included. However, when the tone variables of both local and nonlocal tweets are included in Column (4), the significance of the negative tone in nonlocal tweets disappears and only the negative tone in local tweets has significant predictability for the subsequent three-day returns.

[11] The insignificant results for negative words in news are not consistent with Tetlock (2007) and Tetlock, Saar-Tsechansky, and Macskassy (2008), in which negative words in the Wall Street Journal (WSJ) and Dow Jones News Service (DJNS) predict future returns. This difference can be due to our use of 10 newspapers and business magazines instead of these two specific outlets, our setting missing tone variables as zero for news, the use of different filter (e.g., Tetlock, Saar-Tsechansky, and Macskassy (2008) require at least 5 words that are either positive or negative.), different sample composition (e.g., our choice of six industries), and different sample period. When we repeat our analysis for a smaller sample with non-zero news coverage (N=20,640), we find similar results.

We then examine whether the predictability of negative news in local tweets lasts for longer horizons. These analyses enable us to check whether the relation between the negative tone in local tweets and stock returns is reversed in the subsequent periods, a scenario in which negative tweets by local users moves stock prices away from fundamental values for a short period and then it is followed by stock price reversals. For example, Tetlock (2007) find that media pessimism expressed in the Wall Street Journal predicts downward pressure on market prices and the short-term negative returns are followed by a reversion to fundamentals, suggesting that media contents reflect the demand of noise or liquidity traders rather than value-relevant new information. To examine this possibility, we measure cumulative abnormal returns over several different horizons, including the (t+1, t+5), (t+1, t+10), (t+1, t+30), and (t+1, t+60) windows. Table 6 reports the results. We find that the coefficients on *Local NegTwt* are all significantly negative in Column (1) through Column (4), suggesting no reversal in the subsequent periods. These results provide further support for the view that the negative tone in local tweets contains value-relevant new information about the firm (i.e., information channel) rather than reflecting temporary investor sentiment.

To provide corroborating evidence on the information advantages of local Twitter users, we examine whether the return predicting power of local tweets is pronounced for firms with greater information asymmetry, in which local Twitter users' information advantage over nonlocal users is more likely to be manifested. We use three variables for information asymmetry (size, the number of analyst following, and age) and divide the sample into firms with high and low information asymmetry based on the median value of these variables. We then re-estimate the model to predict the next-day abnormal returns separately for these two groups.

Table 7 reports the regression results for groups with high and low information asymmetry. We find that the predictability of the negative tone in local tweets is present only for firms with high information asymmetry: smaller firms, firms with low analyst following, and younger firms. For example, the coefficient on *Local NegTwt* is significantly negative for smaller firms in Column (1) but is not significant for larger firms in Column (2). Untabulated test confirms that these two coefficients are statistically different between the two groups (*p*-value = 0.09).[12]

Overall, our findings suggest that the negative tone in local tweets predicts future returns, particularly for firms with greater information asymmetry. This return predicting power supports local Twitter users' information advantage about local firms.

Predicting Three-day Abnormal Returns around Earnings Announcement Dates

To provide additional evidence on local Twitter users' ability to predict firms' future prospects, in this section, we investigate whether the fraction of negative words in local tweets predicts subsequent stock returns around earnings announcement dates. We use the earnings announcement as our setting for examination because it represents one of the most information - sensitive events in which local users can exploit their local knowledge. For example, Baik, Kang, and Kim (2010) and Baker, Litov, Wachter, and Wurgler (2010) show that changes in certain types of institutional ownership are related to abnormal returns at the time of subsequent announcement of quarterly earnings, suggesting that those institutions trade based on their private information about future earnings.

[12] The coefficients on *Local NegTwt* between the groups with low and high analyst following and between the groups of young and mature firm are also statistically significant (*p*-value = 0.05 and 0.02, respectively).

To match with the frequency of quarterly earnings, we aggregate Twitter activities between 30 and 3 trading days prior to an earnings announcement date and measure the average ratio of positive and negative words to the total number of words in local and nonlocal tweets during the period (Tetlock, Saar-Tsechansky, and Macskassy, 2008). As the control variables, we include previous quarter's earnings surprise to control for the post-earnings announcement drift effect (Ali, Durtschi, Lev, and Trombley, 2004). We also include the recent stock returns measured over the same period over which Twitter activities are measured (i.e., -30, -3 window) to control for contemporaneous news. In addition, firm size, book-to-market ratios, share turnover, and stock returns two days before the earnings announcement date are also included in the model.

The results in Table 8 show that the negative tone in local tweets predicts subsequent earnings announcement returns, while the tone in nonlocal tweets does not have any predicting power, consistent with our return prediction test results. Similar to our approach in Table 7, we also partition the sample into those with high information asymmetry and those with low information asymmetry based on firm size, the number of analyst following, and age. The regression results are presented in Table 9. We find that the ability of negative news in local tweets to predict subsequent earnings announcement returns is present only for firms with high information asymmetry (smaller firms, firms with low analyst following, and younger firm), consistent with our previous results based on the next-day abnormal returns in Table 7. These results confirm our main findings and further suggest that social media can be a source of value-relevant information about local companies.

5. Additional Tests

To check the robustness of our results and the inference, we conduct several additional tests. Below, we briefly discuss the results of the tests.

Alternative Definitions of Local Users

We use alternative ways to define local users as (1) those located in the same city as the firms' headquarters or (2) those located within 100 km of a firm's headquarters. The results using these alternative definitions of locality are similar to those previously reported. For example, when locality is defined based on same city, the coefficients on *Local NegTwt* is -0.029 (*t*-value = -1.73) and -0.060 (*t*-value=-2.25) for the next-day and next three-day abnormal return regressions, respectively.

Equal-weighted Benchmarks Returns

The results are unchanged when we use equal-weighted Fama-French portfolio returns, rather than value-weighted returns. For example, the coefficient on *Local NegTwt* is -0.025 (*t*-value = -1.77) and -0.057 (*t*-value=-2.31) for the next-day and next three-day abnormal return regressions, respectively.

Inclusion of Non-U.S. Firms

We define local users for non-U.S. firms as those located in the same state as the firms' main branches in the U.S. and then include non-U.S. firms in the sample (N=69,071).We obtain similar results.

Exclusion of Large States

We repeat our analyses after excluding firms whose headquarters are located in California, New York, or Texas as 31% of our sample is from these three states. The results are qualitatively similar.

6. Conclusion

In this paper we examine whether social media activities have any investment value by investigating the return predictability of the tone in messages posted on Twitter. We document that local Twitter users are more likely to send a message about local firms, particularly those firms with high information asymmetry such as firms with low book-to-market, less liquid firms, non-S&P 500 firms, R&D intensive firms, firms with low analyst coverage.

We also find that social media can be a source of private information about local companies. Specifically, we find that whereas the fraction of negative words by local tweets predicts future abnormal returns, no such relation exists between the fraction of negative words by nonlocal tweets and future returns. Moreover, the association between the negative tone in local tweets and future returns is particularly manifested in stocks with high information asymmetry. Finally, we find that the negative tone in local tweets also predicts future earnings announcement returns, supporting the information advantage view of local users. Overall, the findings in this study highlight that social media activities, particularly those of local users, can be used to forecast future stock returns, suggesting an important investment value of social media activities.

Caveats are in order. First, our findings do not directly show whether local individual users earn excess returns from trading. Although it is likely that well-informed local individuals participate in trading to earn profits, we are not able to examine this issue because we do not observe Twitter users' trading behaviors. Second, as we utilize local users' collective view by aggregating all local tweets about a firm on a given day, it is unclear whether the view by each user is value relevant when they are not aggregated.

REFERENCES

Ali, A., Durtschi, C., Lev, B., Trombley, M., 2004. Changes in institutional ownership and subsequent earnings announcement abnormal returns. Journal of Accounting, Auditing & Finance 19, 221-248.

Antweiler, W., Frank, M.Z., 2004. Is all that talk just noise? The information content of internet stock message boards. Journal of Finance 59, 1259-1294.

Bae, K.-H., Stulz, R.M., Tan, H., 2008. Do local analysts know more? A cross-country study of the performance of local analysts and foreign analysts. Journal of Financial Economics 88, 581-606.

Baik, B., Kang, J.-K., Kim, J.-M., 2010. Local institutional investors, information asymmetries, and equity returns. Journal of Financial Economics 97, 81-106.

Baker, M., Litov, L., Wachter, J.A., Wurgler, J., 2010. Can mutual fund managers pick stocks? Evidence from their trades prior to earnings announcements. Journal of Financial and Quantitative Analysis 45, 1111-1131.

Blankespoor, E., Miller, G.S., White, H.D., 2014. The role of dissemination in market liquidity: Evidence from firms' use of Twitter™. The Accounting Review 89, 79-112.

Bollen, J., Mao, H., Zeng, X., 2011. Twitter mood predicts the stock market. Journal of Computational Science 2, 1-8.

Bushee, B.J., Core, J.E., Guay, W., Hamm, S.J.W., 2010. The role of the business press as an information intermediary. Journal of Accounting Research 48, 1-19.

Chen, H., Hwang, B., Liu, B., 2013. The economic consequences of having social executives. Working paper, Purdue University.

Chen, H., De, P., Hu, Y., Hwang, B.-H., 2014. Wisdom of crowds: The value of stock opinions transmitted through social media. Review of Financial Studies 27, 1367-1403.

CNN, 2012. Titter unveils ‘cashtags’ to track stock symbols. July 31, 2012. available at http://money.cnn.com/2012/07/31/technology/twitter-cashtag/

Constantinides, E., Romero, C., Boria, M., 2009. Social media: A new frontier for retailers? European Retail Research, 1-28.

Coval, J.D., Moskowitz, T.J., 1999. Home bias at home: Local equity preference in domestic portfolios. Journal of Finance 54, 2045-2073.

Coval, J.D., Moskowitz, T. J., 2001. The geography of investment: Informed trading and asset prices. Journal of Political Economy 109, 811-841.

Culnan, M., Patrick, J., McHugh, J., Zubillaga, J., 2010. How large US companies can use Twitter and other social media to gain business value. MIS Quarterly Executive 9, 243-259.

Das, S. R., Chen, M. Y., 2007. Yahoo! For Amazon: Sentiment extraction from small talk on the web. Management Science 53, 1375-1388.

Fama, E.F., French, K.R., 1996. Multifactor explanations of asset pricing anomalies. Journal of Finance 51, 55-84.

Financial Times, 2011. Twitter research promises trading success. May 8, 2011. Available at http://www.ft.com/cms/s/0/fd34524a-782c-11e0-b90e-00144feabdc0.html#axzz3686OpMOq

Giannini, R., Irvine, P., Shu, T., 2013. Do local investors know more? A direct examination of individual investors' information set. Working paper. BlueCrest Capital Management, Texas Christian University, and University of Georgia.

Green, J., Kesselheim, S., 2010. Pharmaceutical marketing and the new social media. New England Journal of Medicine 363, 2087-2089.

Ivkovic, Z., Weisbenner, S., 2005. Local does as local is: Information content of the geography of individual investors' common stock investments. Journal of Finance 60, 267-306.

Ivkovic, Z., Weisbenner, S., 2007. Information diffusion effects in individual investors' common stock purchases: Covet thy neighbors' investment choices. Review of Financial Studies 20, 1327-1357.

Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. Journal of Finance 48, 65-91.

Kim, Y. H., Meschke, F., 2011. CEO interviews on CNBC. Working paper. Nanyang Technological University and University of Kansas.

Kumar, S., Morstatter, F., Liu, H., 2013. Twitter data analytics. Springer.

Loughran, T.I.M., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance 66, 35-65.

Noone, B., McGuire, K., Rohlfs, K., 2011. Social media meets hotel revenue management: Opportunities, issues, and unanswered questions. Journal of Revenue and Pricing Management 10, 293-305.

Malloy, C.J., 2005. The geography of equity analysis. Journal of Finance 60, 719-755.

Phil, K., Nguyen, B., 2013. Exploring the role of the online customer experience in firms' multi-channel strategy: An empirical analysis of the retail banking services sector. Journal of Strategic Marketing 21, 429-442.

Rock, S., Sedo, S., Willenborg, M., 2000. Analyst following and count-data econometrics. Journal of Accounting and Economics 30, 351-373.

Seasholes, M.S., Zhu, N., 2010. Individual investors and local bias. Journal of Finance 65, 1987-2010.

Securities and Exchange Commission (SEC). 2013. SEC says social media OK for company announcements if investors are alerted. available at http://www.sec.gov/News/PressRelease/Detail/PressRelease/1365171513574#.U6weBvldX6w

Sprenger, T.O., Tumasjan, A., Sandner, P.G., Welpe, I.M., 2013. Tweets and trades: The information content of stock microblogs. European Financial Management, forthcoming.

Takhteyev, Y., Gruzd, A., Wellman, B., 2012. Geography of Twitter networks. Social Networks 34, 73-81.

Tetlock, P.C., 2007. Giving content to investor sentiment: The role of media in the stock market. Journal of Finance 62, 1139-1168.

Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S., 2008. More than words: Quantifying language to measure firms' fundamentals. Journal of Finance 63, 1437-1467.

Wang, G., Wang, T., Wang, B., Sambasivan, D., Zhang, Z., Zheng, H., Zhao, B. Y. 2014. Crowds on Wall street: Extracting value from social investing platforms. Working paper.

Wysocki, P., 1998. Cheap talk on the web: The determinants of postings on stock message boards. Working paper, University of Michigan.

Yu, Y., Duan, W., Cao, Q., 2013. A Dynamic model of the impact of social media and conventional media on firm performance. Decision Support Systems 55, 919-926.

Appendix I– Twitter Data Gathering Process

As Figure 1 illustrates, our data gathering process consists of four steps.

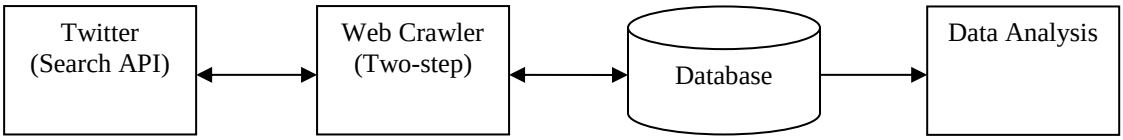

Figure 1 - Data Gathering Procedure

1) Twitter API selection

We use the Twitter Search application programming interface (API) to conduct a city location (geo-code) search and a general term (text)-based search (i.e., company name).[13]

2) Web Crawler

We designed a web crawler program using the python scripts (codes) to fetch location-based company specific tweets from the Twitter Search API. The web crawler consists of two steps: location-based search and text-based search. For location-based search, we first define locality as a city measured as the area within a specific radius from the geo-code (latitude and longitude) of the city center. For example, for Houston, its radius is set as 20 km to cover the area (1,560 km^2) and then geo-code=(29.7805°N 95.3863°W, 20km) presents Houston and its surrendering area. To obtain the data for the state level, data of cites belonging to the same state are aggregated. Next, text-based search using company name is employed to collect company specific tweets in the abovementioned city from Twitter API. In our data collection, top 293 cities based on the 2010 United States Census and 1,044 public firms were employed to gather the twitter data for a period of nine months from July 2011 to March 2012. Detailed data files (categories) collected by the Twitter Search API include user location (self-reported), time of tweet, user id, user name, tweet content (text), geo (if mobile), and sources used to tweet.

3) Database

We created a database to store the twitter data collected by our web crawler. The python scripts requested data for different locations and companies from the Twitter Search API sequentially. The results were then parsed and stored in the database. The database is used to store and retrieve query results. In addition, the database rejects duplicate tweets by using tweet id as a unique key.

4) Data Analysis

To capture the individual users' view expressed in tweets, we used a "bags of words" approach. In other words, we counted negative and positive words defined by Loughran and McDonald (2011) and normalize them by the number of total words.

[13] There are other types of Twitter APIs that can be used in Twitter data collection. For example, the REST API can provide us with core Twitter data, such as user profile and user list. However, its search capability is limited and the location information provided from the user profile might not be the actual location of the user (Kumar, Morstatter, and Liu, 2013). The streaming API, other API available to us, provides a near real-time high-volume access to Twitter data. However, it is not efficient because it only allows one connection and one set of filters at a time (i.e., to change filters, reconnection is required).

Appendix II - Samples of Tweets by Local and Nonlocal Users

This appendix provides examples of two cases in which local and nonlocal Twitter users refer to the same event relating the same firm but only local users use negative words in tweets. Negative words identified by the word list by Loughran and McDonald (2011) are underlined.

Case 1

(1) Tweet by local user

"*The xxx bank has site &mobile access problem and we are unable to do normal transactions online. It makes our life so difficult."*

(2) Tweet by nonlocal user

"*We're seeing high traffic causing online & mobile access issues at the bank, like a cyber traffic jam*."

Case 2

(1) Tweet by local user

"*Q1 earnings: substantial decline in revenue expected due to recent patent expirations and new drug failure."*

(2) Tweet by nonlocal user

"*CEO on Q1 earnings: earning is down but patent expirations partially offset by growth in other products.*"

Appendix III- Variable Definitions

Variable	Definition
CountTwt	The total number of tweets about the firm on a given day.
Local CountTwt (*Nonlocal CountTwt*)	The total number of local (nonlocal) tweets about the firm on a given day.
PosTwt (*NegTwt*)	The ratio of the number of positive (negative) words to the total number of words in all tweets about the firm on a given day. The word list for positive (negative) words is from Loughran and McDonald (2011).
Local PosTwt (*Local NegTwt*)	The ratio of the number of positive (negative) words to the total number of words in all local tweets about the firm on a given day. Local tweets are defined as those by local users who are located in the same state as the firm's headquarters.
Nonlocal PosTwt (*Nonlocal NegTwt*)	The ratio of the number of positive (negative) words to the total number of words in all nonlocal tweets about the firm on a given day. Nonlocal tweets are defined as those by nonlocal users who are located in the different state as the firm's headquarters.
CountNews	The total number of news about the firm on a given day.
Local CountNews (*Nonlocal CountNews*)	The total number of local (nonlocal) news about the firm on a given day.
PosNews (*NegNews*)	The ratio of the number of positive (negative) words to the total number of words in the newspapers and business magazines about the firm on a given day. We select 10 outlets including ABC News, New York Times, Reuters, USA Today, Fox News, Wall Street Journal, Washington Post, CNN, Economist, and Forbes. The word list for positive (negative) words is from Loughran and McDonald (2011).
Local PosNews (Local NegNews)	The ratio of the number of positive (negative) words to the total number of words in all local news about the firm on a given day. Local news are defined as those whose headquarters are located in the same state as the firm's headquarters.
Nonlocal PosNews (Nonlocal NegNews)	The ratio of the number of positive (negative) words to the total number of words in all nonlocal news about the firm on a given day. Nonlocal news are defined as those whose headquarters are located in the different states as the firm's headquarters.
AR (t+n)	The firm's abnormal returns on day *t*+n, where *t* is the day of tweets or the ensuing trading day if the tweet is on a non-trading day. Abnormal returns are calculated as raw returns adjusted for 25 (5*5) value-weighted portfolios by size and book-to-market (B/M). The daily returns and breakpoints of the 25 size-B/M portfolios are from Kenneth French's web site.
AR(t+m,t+n)	The firm's cumulative abnormal returns from *t*+m to *t*+n, where *t* is the day of tweets or the ensuing trading day if the tweet is on a non-trading day. Abnormal returns are calculated as raw returns adjusted for 25 (5*5) value-weighted portfolios by size and book-to-market (B/M).
SUE	The firm's standardized unexpected quarterly earnings measured as the seasonal difference in quarterly earnings per share scaled by the end of quarter price.
Size	Log of market capitalization at the end of the preceding quarter.

BM	Book-to-market ratio at the end of the preceding quarter.
Turnover	The average of monthly trading volume divided by the number of shares outstanding over the 12-month period ending at the end of the preceding quarter.
Ret Volatility	Stock return volatility measured as the standard deviation of monthly returns over the 12-month period ending at the end of the preceding quarter.
Price	Per share stock price at the end of the preceding quarter.
S&P500	An indicator variable that takes the value of 1 if the firm is included in the S&P 500 index, and 0 otherwise.
Age	Log of the number of months since a firm's first stock return appears in CRSP
Dividend yield	Cash dividends in the preceding quarter divided by share price at the end of the preceding quarter
R&D	Research and development expenses (zero for missing values) divided by total assets; and
AnalystD	An indicator variable that takes the value of 1 if the number of analysts following the firm is above the median, and 0 otherwise.

Table 1 Sample Selection Process

	The number of firm-day level observations	The number of unique firms
Social media coverage data (July 1, 2011 – March 31, 2012)	287,085	1,044
Merge with Compustat and CRSP	228,362	855
Exclude non-U.S. firms	200,765	752
Require dependent variables and control variables	165,893	654
Require non-missing Twitter coverage (Data for the return prediction test)	63,103	646

This table shows our sample selection procedure for the return prediction test.

Table 2 Descriptive Statistics

Variables	N	Mean	Standard Deviation	Q1	Median	Q3
CountTwt	63,103	8.748	7.858	5	6	11
Local CountTwt	63,103	5.667	5.387	3	4	7
Nonlocal CountTwt	63,103	3.081	3.001	1	2	4
PosTwt	63,103	0.002	0.007	0	0	0
NegTwt	63,103	0.002	0.008	0	0	0
Local PosTwt	63,103	0.002	0.008	0	0	0
Local NegTwt	63,103	0.002	0.010	0	0	0
Nonlocal PosTwt	63,103	0.002	0.009	0	0	0
Nonlocal NegTwt	63,103	0.002	0.011	0	0	0
CountNews	63,103	0.620	1.058	0	0	1.000
Local CountNews	63,103	0.441	0.752	0	0	1.000
Nonlocal CountNews	63,103	0.178	0.426	0	0	0.000
PosNews	20,640	0.002	0.003	0	0	0.003
NegNews	20,640	0.006	0.008	0	0	0.010
Local PosNews	20,033	0.002	0.004	0	0	0.004
Local NegNews	20,033	0.006	0.009	0	0	0.010
Nonlocal PosNews	10,299	0.002	0.005	0	0	0.000
Nonlocal NegNews	10,299	0.008	0.012	0	0	0.012
AR (t+1)	63,103	0.000	0.031	-0.012	0.000	0.011
AR(t +1, t +3)	63,103	0.001	0.049	-0.019	0.000	0.020
AR(t +1, t +5)	63,103	0.001	0.063	-0.025	0.001	0.026
AR(t +1, t +10)	63,103	0.003	0.088	-0.034	0.002	0.038
AR(t +1, t +30)	63,103	0.009	0.148	-0.056	0.006	0.068
AR(t +1, t +60)	63,103	0.028	0.210	-0.074	0.018	0.116
AR (t +0)	63,103	0.001	0.031	-0.012	0.000	0.012
AR(t -1)	63,103	0.001	0.030	-0.011	0.000	0.011
AR(t -2)	63,103	0.001	0.031	-0.011	0.000	0.012
AR(t -30, t -3)	63,103	0.011	0.148	-0.058	0.008	0.075
SUE	63,103	-0.004	0.114	-0.005	0.002	0.007
Size	63,103	6.412	2.191	4.793	6.309	7.905
BM	63,103	0.768	0.725	0.284	0.545	1.033
Turnover	63,103	0.192	0.160	0.063	0.155	0.277

This table reports descriptive statistics for the variables used in our main analyses. See Appendix III for variable definitions.

Table 3 Correlations

	NegTwt	*Local PosTwt*	*Local NegTwt*	*Nonlocal PosTwt*	*Nonlocal NegTwt*	*AR(t+1)*	*AR(t+1, t+3)*	*AR(t+0)*	*AR(t -1)*	*AR(t -2)*	*AR(t-30, t-3)*	*SUE*	*Size*	*BM*	*Turnover*
PosTwt	**0.067**	**0.939**	**0.071**	**0.784**	**0.039**	-0.006	0.000	-0.004	0.001	**0.012**	-0.005	**0.043**	**-0.043**	**-0.009**	**-0.015**
NegTwt		**0.062**	**0.936**	**0.053**	**0.765**	**-0.008**	**-0.013**	-0.007	-0.004	**0.012**	0.002	**0.021**	**0.026**	**0.023**	**-0.012**
Local PosTwt			**0.070**	**0.551**	**0.030**	-0.007	-0.001	-0.004	0.003	**0.012**	-0.003	**0.040**	**-0.039**	**-0.010**	**-0.015**
Local NegTwt				**0.050**	**0.520**	**-0.010**	**-0.013**	-0.005	-0.004	**0.014**	0.001	**0.020**	**0.026**	**0.022**	**-0.012**
Nonlocal PosTwt					**0.038**	-0.001	0.001	-0.006	-0.002	0.007	**-0.008**	**0.036**	**-0.036**	-0.003	**-0.012**
Nonlocal NegTwt						-0.003	**-0.008**	-0.005	-0.001	0.005	0.004	**0.015**	**0.021**	**0.016**	**-0.008**
AR(t+1)							**0.564**	**-0.088**	**-0.025**	**-0.036**	-0.004	**0.016**	0.004	**0.009**	-0.004
AR(t+1, t +3)								**-0.071**	**-0.031**	**-0.019**	0.005	**0.009**	**-0.012**	**0.020**	**-0.017**
AR(t+0)									**-0.079**	-0.003	-0.002	-0.006	-0.007	0.007	-0.006
AR(t -1)										**-0.060**	-0.002	0.002	**-0.020**	**0.016**	**-0.012**
AR(t -2)											-0.005	0.003	**-0.015**	**0.011**	**-0.013**
AR(t-30, t-3)												**0.008**	**-0.027**	**0.017**	0.000
SUE													**0.040**	**-0.067**	**-0.056**
Size														**-0.509**	**0.431**
BM															**-0.302**

This table presents the Pearson correlation coefficients. Bold values are significant at the 5% level or better (two-tailed). The sample consists of 63,103 observations for which the Twitter tone variables and other key variables are available. See Appendix III for variable definitions.

Table 4 Determinants of Twitter Coverage and Local Twitter Coverage

Dependent variable =	(1) An indicator for daily Twitter coverage	(2) Daily count of Twitter coverage	(3) An indicator for local Twitter coverage	(4) The ratio of local tweet count to total tweet count
Intercept	-1.576***	0.250***	-2.258***	0.677
	(0.00)	(0.00)	(0.00)	(0.00)***
Size	0.051***	0.044***	0.020	0.000
	(0.00)	(0.00)	(0.17)	(0.62)
BM	0.183***	0.164***	-0.258***	-0.004***
	(0.00)	(0.00)	(0.00)	(0.00)
Ret volatility	0.524***	0.542***	0.136	-0.003
	(0.00)	(0.00)	(0.67)	(0.81)
Turnover	1.125**	1.103***	-1.116***	-0.009*
	(0.02)	(0.00)	(0.00)	(0.08)
Price	0.001***	0.001***	0.002*	0.000**
	(0.00)	(0.00)	(0.08)	(0.03)
S&P500	0.140**	0.186***	-0.316***	-0.004*
	(0.01)	(0.00)	(0.00)	(0.08)
AR(t -1)	0.770***	0.483*	0.780	-0.023
	(0.01)	(0.06)	(0.29)	(0.32)
AR(t -2)	0.916***	0.675***	1.026	0.025
	(0.00)	(0.01)	(0.22)	(0.34)
AR(t -30, t -3)	0.239***	0.179***	-0.378***	-0.012**
	(0.00)	(0.00)	(0.00)	(0.01)
Age	0.064***	0.028***	-0.023	-0.002*
	(0.00)	(0.00)	(0.18)	(0.04)
Dividend yield	8.036***	5.542***	-17.967***	-0.148***
	(0.00)	(0.00)	(0.00)	(0.00)
R&D	-2.029***	-4.129***	2.087***	0.017
	(0.00)	(0.00)	(0.00)	(0.46)
AnalystD	0.215***	0.102***	-0.117***	-0.004*
	(0.00)	(0.00)	(0.01)	(0.06)
Pseudo R^2	2.46%	0.60%	1.55%	-0.20%
N	169,007	169,007	68,233	68,233

This table reports the results on the determinants of Twitter coverage (Columns 1 and 2) and local Twitter coverage compared to nonlocal Twitter coverage (Columns 3 and 4). Column (1) reports the logistic regression estimates of an indicator variable for daily Twitter coverage on the determinants of Twitter coverage. The dependent variable is an indicator variable that takes the value of 1 if the firm is covered in Twitter for a given day, and 0 otherwise. Column (2) reports the negative binomial regression estimates of daily count of Twitter coverage on the determinants of coverage. The dependent variable is the number of tweets that mention the firm for a given day. In Columns (1) and (2), the sample is 169,007 firm-day observations during the July 2011-March 2012 period for which the determinant variables are available. Column (3) reports the logit regression estimates of an indicator variable for local tweet coverage. The dependent variable has a value of 1 if the firm is covered by local tweets but is not covered by nonlocal tweets, and 0 otherwise (i.e., covered by both local and nonlocal tweets or covered only by nonlocal tweets). Column (4) reports the tobit regression estimates of the ratio of local tweets on the determinants of Twitter coverage. The dependent variable is the ratio of the number of local tweets to the number of total tweets that mention the firm on a given day. The sample is 68,233 firm-day observations for which the Twitter coverage variables and determinant variables are available. In Columns (1)-(4), the *p*-values in parentheses are based on robust standard errors clustered by date. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels, respectively, in two-tailed tests.

Table 5 Predicting Returns using Positive and Negative Words in Local Tweets on Twitter

Panel A. Predicting (*t*+1) Returns using Positive and Negative Words in Local Tweets

	Dependent variable = AR(t+1)			
	(1)	(2)	(3)	(4)
Intercept	-0.001	-0.001	-0.001	-0.001
	(-0.45)	(-0.45)	(-0.48)	(-0.45)
PosTwt	-0.022			
	(-0.67)			
NegTwt	-0.035*			
	(-1.85)			
Local PosTwt		-0.023		-0.028
		(-0.81)		(-1.02)
Local NegTwt		-0.033**		-0.035**
		(-2.08)		(-2.06)
Nonlocal PosTwt			-0.006	0.009
			(-0.25)	(0.42)
Nonlocal NegTwt			-0.012	0.004
			(-0.90)	(0.27)
PosNews	-0.029			
	(-0.29)			
NegNews	0.021			
	(0.64)			
Local PosNews		-0.032		-0.031
		(-0.37)		(-0.33)
Local NegNews		0.018		0.011
		(0.51)		(0.29)
Nonlocal PosNews			-0.023	-0.012
			(-0.35)	(-0.19)
Nonlocal NegNews			0.018	0.015
			(0.67)	(0.54)
AR(t +0)	-0.089***	-0.089***	-0.089***	-0.089***
	(-3.83)	(-3.83)	(-3.82)	(-3.83)
AR(t -1)	-0.035*	-0.035*	-0.035**	-0.035**
	(-1.92)	(-1.92)	(-1.92)	(-1.92)
AR(t -2)	-0.038***	-0.038***	-0.038***	-0.038***
	(-2.62)	(-2.62)	(-2.63)	(-2.62)
AR(t -30, t -3)	-0.001	-0.001	-0.001	-0.001
	(-0.37)	(-0.37)	(-0.37)	(-0.37)
SUE	0.004	0.004	0.004	0.004
	(1.48)	(1.48)	(1.46)	(1.48)
Size	0.000	0.000	0.000	0.000
	(0.69)	(0.69)	(0.69)	(0.69)
BM	0.001	0.001	0.001	0.001
	(1.56)	(1.56)	(1.55)	(1.56)
Turnover	-0.001	-0.001	-0.001	-0.001
	(-0.41)	(-0.41)	(-0.40)	(-0.41)
Adjusted R^2	1.06%	1.06%	1.05%	1.05%
N	63,103	63,103	63,103	63,103

Table 5 (Continued)

Panel B. Predicting (*t*+1, *t*+3) Returns using Positive and Negative Words in Local Tweets

	Dependent variable = AR(t+1, t+3)			
	(1)	(2)	(3)	(4)
Intercept	0.001	0.001	0.001	0.001
	(0.27)	(0.28)	(0.26)	(0.28)
PosTwt	0.002			
	(0.05)			
NegTwt	-0.086***			
	(-2.98)			
Local PosTwt		0.000		-0.005
		(0.00)		(-0.14)
Local NegTwt		-0.074***		-0.068**
		(-2.77)		(-2.24)
Nonlocal PosTwt			0.005	0.009
			(0.16)	(0.31)
Nonlocal NegTwt			-0.041**	-0.010
			(-2.37)	(-0.51)
PosNews	0.002			
	(0.02)			
NegNews	0.020			
	(0.36)			
Local PosNews		0.026		0.056
		(0.22)		(0.45)
Local NegNews		0.005		-0.022
		(0.09)		(-0.39)
Nonlocal PosNews			-0.096	-0.111
			(-0.98)	(-1.02)
Nonlocal NegNews			0.053	0.061
			(1.14)	(1.25)
AR(t +0)	-0.116**	-0.116***	-0.116***	-0.116***
	(-3.23)	(-3.23)	(-3.23)	(-3.23)
AR(t -1)	-0.064**	-0.064**	-0.064**	-0.064**
	(-2.52)	(-2.52)	(-2.52)	(-2.52)
AR(t -2)	-0.035	-0.035	-0.035	-0.035
	(-1.61)	(-1.60)	(-1.62)	(-1.60)
AR(t -30, t -3)	0.002	0.002	0.002	0.002
	(0.40)	(0.40)	(0.40)	(0.40)
SUE	0.004	0.004	0.004	0.004
	(1.18)	(1.18)	(1.16)	(1.18)
Size	0.000	0.000	0.000	0.000
	(0.14)	(0.14)	(0.11)	(0.14)
BM	0.001	0.001	0.001	0.001
	(1.45)	(1.45)	(1.43)	(1.45)
Turnover	-0.004	-0.004	-0.004	-0.004
	(-0.82)	(-0.82)	(-0.81)	(-0.82)
Adjusted R^2	0.75%	0.75%	0.74%	0.75%
N	63,103	63,103	63,103	63,103

This table presents the regression estimates of firms' future returns on positive and negative words in tweets. In Panel A, the dependent variable is the firm's abnormal returns on day *t*+1, where *t* is the day of tweets or the ensuing trading day if the tweet is on a non-trading day. Abnormal returns are calculated as raw returns adjusted for 25 (5*5) value-weighted portfolios by size and book-to-market (B/M). In Panel B, the dependent variable is three-day cumulative abnormal returns from *t*+1 to *t*+3. *PosTwt* (*NegNews*) is the ratio of the number of positive (negative) words to the total number of words in all tweets about the firm on a given day. *Local PosTwt* (*Local NegTwt)* and *Nonlocal PosTwt* (*Nonlocal NegTwt*) are the ratio of the number of positive (negative) words to

the total number of words in all local and nonlocal tweets about the firm on a given day. See Appendix III for other variable definitions. The sample consists of 63,103 observations for which the Twitter tone variables and other variables are available. The *t*-values in parentheses are based on robust standard errors clustered by date. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels, respectively, in two-tailed tests.

Table 6 Predicting Returns using Positive and Negative Words in Local Tweets on Twitter: Longer Horizons

Dependent variable =	*AR(t +1, t +5)*	*AR(t +1, t +10)*	*AR(t +1, t +30)*	*AR(t +1, t +60)*
	(1)	(2)	(3)	(4)
Intercept	-0.002	-0.001	-0.001	0.052
	(-0.49)	(-0.29)	(-0.24)	(6.02)
Local PosTwt	0.001	0.016	0.410	0.266
	(0.02)	(0.23)	(2.55)	(1.36)
Local NegTwt	-0.132***	-0.135***	-0.271***	-0.309***
	(-4.23)	(-3.09)	(-3.53)	(-2.81)
Nonlocal PosTwt	-0.029	-0.042	0.249*	0.303
	(-0.79)	(-0.79)	(1.70)	(1.53)
Nonlocal NegTwt	0.003	-0.037	0.027	-0.018
	(0.09)	(-0.93)	(0.50)	(-0.25)
Local PosNews	0.021	0.143	-0.002	0.332
	(0.13)	(0.65)	(-0.01)	(0.64)
Local NegNews	0.094	-0.037	-0.209	-0.055
	(1.34)	(-0.38)	(-1.34)	(-0.24)
Nonlocal PosNews	-0.190	-0.176	0.326	0.542
	(-1.46)	(-0.93)	(0.72)	(0.88)
Nonlocal NegNews	0.090	0.072	0.127	0.471**
	(1.51)	(1.10)	(0.91)	(2.26)
AR(t +0)	-0.103**	-0.146***	-0.171**	-0.188**
	(-2.17)	(-2.83)	(-2.47)	(-2.20)
AR(t -1)	-0.094***	-0.149***	-0.118***	-0.194***
	(-2.95)	(-4.20)	(-2.66)	(-3.56)
AR(t -2)	-0.076***	-0.082***	-0.090***	-0.182***
	(-2.71)	(-3.01)	(-2.65)	(-3.55)
AR(t -30 t,-3)	-0.003	0.001	-0.002	-0.087***
	(-0.59)	(0.11)	(-0.14)	(-4.82)
SUE	0.003	0.001	-0.015	-0.011
	(0.82)	(0.23)	(-1.37)	(-0.55)
Size	0.000	0.000	0.001	-0.004***
	(1.25)	(1.32)	(1.17)	(-4.87)
BM	0.002**	0.004**	0.008***	0.003
	(2.27)	(2.21)	(3.12)	(0.89)
Turnover	-0.004	-0.002	0.000	-0.005
	(-0.67)	(-0.29)	(0.04)	(-0.32)
Adjusted R^2	0.66%	0.62%	0.46%	0.89%
N	63,103	63,103	63,103	63,103

This table presents the regression estimates of firms' future returns on positive and negative words in tweets on Twitter. In Column (1), the dependent variable is the firm's cumulative abnormal returns from *t*+1 to *t*+5, where *t* is the day of tweets or the ensuing trading day if the tweet is on a non-trading day. Abnormal returns are calculated as raw returns adjusted for 25 (5*5) value-weighted portfolios by size and book-to-market (B/M). In Columns (2), (3), and (4), the return cumulation period is (+1, +10), (+1, +30), and (+1, +60), respectively. *PosTwt* (*NegNews*) is the ratio of the number of positive (negative) words to the total number of words in all tweets about the firm on a given day. *Local PosTwt* (*Local NegTwt)* and *Nonlocal PosTwt* (*Nonlocal NegTwt*) are the ratio of the number of positive (negative) words to the total number of words in all local and nonlocal tweets about the firm on a given day. See Appendix III for other variable definitions. The sample consists of 63,103 observations for which the Twitter tone variables and other variables are available. The *t*-values in parentheses are based on robust standard errors clustered by date. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels, respectively, in two-tailed tests.

Table 7 Predicting Returns using Positive and Negative Words in Local Tweets on Twitter: High and Low Information Asymmetry

	Dependent variable =AR(t +1)					
Partitioning variable	Size		Analyst followings		Age	
	Small	Large	Low	High	Young	Mature
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	-0.001	0.002	-0.002	0.001	-0.001	-0.002
	(-0.11)	(1.40)	(-0.33)	(1.21)	(-0.28)	(-0.60)
Local PosTwt	-0.022	-0.036	-0.021	-0.040	0.024	-0.084**
	(-0.51)	(-1.29)	(-0.47)	(-1.23)	(0.80)	(-2.10)
Local NegTwt	-0.059**	-0.002	-0.072**	0.004	-0.067**	-0.004
	(-1.99)	(-0.14)	(-2.18)	(0.25)	(-2.56)	(-0.22)
Nonlocal PosTwt	0.005	0.016	0.017	-0.005	-0.025	0.038
	(0.15)	(0.74)	(0.56)	(-0.19)	(-1.06)	(1.24)
Nonlocal NegTwt	0.010	0.001	0.015	-0.008	0.022	-0.014
	(0.38)	(0.12)	(0.55)	(-0.66)	(1.05)	(-0.78)
Local PosNews	-0.098	0.033	-0.065	-0.008	0.022	-0.062
	(-0.67)	(0.38)	(-0.44)	(-0.10)	(0.19)	(-0.55)
Local NegNews	0.008	0.021	-0.034	0.066*	0.007	0.015
	(0.12)	(0.52)	(-0.57)	(1.89)	(0.14)	(0.31)
Nonlocal PosNews	-0.042	0.011	-0.002	-0.021	-0.042	0.017
	(-0.40)	(0.17)	(-0.02)	(-0.30)	(-0.49)	(0.20)
Nonlocal NegNews	0.043	-0.019	0.042	-0.017	0.039	-0.007
	(0.92)	(-0.77)	(0.88)	(-0.57)	(0.99)	(-0.20)
AR(t +0)	-0.110***	-0.011	-0.116***	-0.022	-0.056**	-0.137***
	(-3.52)	(-0.55)	(-3.37)	(-1.15)	(-2.21)	(-5.46)
AR(t -1)	-0.041*	-0.026	-0.046*	-0.021	-0.035*	-0.041*
	(-1.94)	(-1.63)	(-1.95)	(-1.44)	(-1.84)	(-1.66)
AR(t -2)	-0.044***	-0.024	-0.049***	-0.017	-0.037**	-0.039**
	(-2.66)	(-1.40)	(-2.90)	(-0.95)	(-2.29)	(-1.98)
AR(t -30, t -3)	-0.001	-0.001	-0.001	0.000	0.000	-0.003
	(-0.33)	(-0.51)	(-0.50)	(0.07)	(-0.10)	(-0.94)
SUE	0.005	0.001	0.007*	0.001	-0.001	0.009***
	(1.43)	(0.19)	(1.66)	(0.32)	(-0.18)	(2.62)
Size	0.000	0.000	0.000	0.000	0.000	0.000
	(0.04)	(-0.93)	(0.23)	(-0.25)	(0.38)	(0.88)
BM	0.001	0.000	0.001	-0.001	0.001	0.001
	(1.26)	(-0.22)	(1.60)	(-1.67)*	(1.05)	(1.63)
Turnover	-0.002	-0.001	-0.001	-0.002	-0.001	-0.001
	(-0.58)	(-0.80)	(-0.15)	(-1.19)	(-0.27)	(-0.56)
Adjusted R^2	1.51%	0.13%	1.71%	0.15%	0.56%	2.14%
N	31,551	31,551	31,938	31,165	31,508	31,595

This table presents the regression estimates of firms' t+1 returns on positive and negative words in tweets on Twitter by the extent of information asymmetry. The sample is divided into those with high information asymmetry and those with low information asymmetry based on the sample median of each information asymmetry variable (i.e., firm size, the number of analyst following, and firm age). The dependent variable is the firm's abnormal returns on day *t*+1, where *t* is the day of tweets or the ensuing trading day if the tweet is on a non-trading day. Abnormal returns are calculated as raw returns adjusted for 25 (5*5) value-weighted portfolios by size and book-to-market (B/M). *PosTwt* (*NegNews*) is the ratio of the number of positive (negative) words to the total number of words in all tweets about the firm on a given day. *Local PosTwt* (*Local NegTwt)* and *Nonlocal PosTwt* (*Nonlocal NegTwt*) are the ratio of the number of positive (negative) words to the total number of words in all local and nonlocal tweets about the firm on a given day. See Appendix III for other variable definitions. The sample consists of 63,103 observations for which the Twitter tone variables and other variables are available. The *t*-values in parentheses are based on robust standard errors clustered by date. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels, respectively, in two-tailed tests.

Table 8 Predicting Three-day Abnormal Returns around Earnings Announcement Dates

	Dependent variable = $AR_{(t-1, t+1)}$ around earnings announcement dates			
	(1)	(2)	(3)	(4)
Intercept	-0.002	-0.002	-0.002	-0.002
	(-0.30)	(-0.34)	(-0.40)	(-0.30)
$PosTwt_{-30,-3}$	-0.267			
	(-0.41)			
$NegTwt_{-30,-3}$	-0.573			
	(-1.82)			
Local $PosTwt_{-30,-3}$		-0.191		-0.169
		(-0.33)		(-0.29)
Local $NegTwt_{-30,-3}$		-0.543*		-0.525**
		(-2.03)		(-2.52)
Nonlocal $PosTwt_{-30,-3}$			-0.147	-0.023
			(-0.38)	(-0.21)
Nonlocal $NegTwt_{-30,-3}$			-0.305	-0.006
			(-1.36)	(-0.04)
$PosNews_{-30,-3}$	0.340			
	(0.19)			
$NegNews_{-30,-3}$	-0.012			
	(-0.03)			
Local $PosNews_{-30,-3}$		0.195		0.694
		(0.13)		(0.40)
Local $NegNews_{-30,-3}$		0.032		0.092
		(0.09)		(0.24)
Nonlocal $PosNews_{-30,-3}$			-1.173	-1.473
			(-0.98)	(-0.98)
Nonlocal $NegNews_{-30,-3}$			-0.015	-0.050
			(-0.04)	(-0.11)
Lagged SUE	-0.019	-0.019	-0.020	-0.018
	(-0.61)	(-0.61)	(-0.66)	(-0.60)
Size	0.000	0.000	0.001	0.001
	(0.47)	(0.51)	(0.54)	(0.54)
BM	0.004	0.004	0.005	0.004
	(0.83)	(0.85)	(0.83)	(0.84)
Turnover	-0.004	-0.004	-0.004	-0.005
	(-0.19)	(-0.19)	(-0.17)	(-0.22)
$AR_{(t-30, t-3)}$	0.003	0.003	0.003	0.003
	(0.17)	(0.17)	(0.18)	(0.18)
$AR_{(t-2)}$	-0.274**	-0.272**	-0.279*	-0.273**
	(-2.38)	(-2.38)	(-2.34)	(-2.37)
Adjusted R^2	0.07%	0.05%	0.03%	-0.21%
N	1,197	1,197	1,197	1,197

This table presents the regression estimates of firms' subsequent three-day abnormal returns around quarterly earnings announcement dates on positive and negative words in tweets on Twitter. The dependent variable is the firm's three-day (-1, +1) cumulative abnormal returns centered on the subsequent quarterly earnings announcement date (day *t*). Abnormal returns are calculated as raw returns adjusted for 25 (5*5) value-weighted portfolios by size and book-to-market (B/M). *$PosTwt_{-30,-3}$* (*$NegNews_{-30,-3}$*) is the average ratio of positive (negative) words to the total number of words between 30 and 3 trading days prior to an earnings announcement date. *Local $PosTwt_{-30,-3}$* (*Local $NegTwt_{-30,-3}$)* and *Nonlocal $PosTwt_{-30,-3}$* (*Nonlocal $NegTwt_{-30,-3}$*) are the average ratio of positive (negative) words to the total number of words in all local and nonlocal tweets about the firm between 30 and 3 trading days prior to an earnings announcement date. The news variables (*$PosNews_{-30,-3}$*, *$NegNews_{-30,-3}$*, *Local $PosNews_{-30,-3}$*, *Local $NegNews_{-30,-3}$*, *Nonlocal $PosNews_{-30,-3}$*, and *Nonlocal $NegNews_{-30,-3}$*) are similarly defined. *Lagged SUE* is the firm's standardized unexpected quarterly earnings in the previous quarter. It is measured as the seasonal difference in quarterly earnings per share scaled by the end of quarter price. *$AR_{(t-}$*

$_{30, t-3)}$ is cumulative abnormal returns from *t*-30 to *t*-3, where *t* is the earnings announcement date. $AR_{(t-2)}$ is abnormal returns on day *t*-2, where *t* is the earnings announcement date. See Appendix III for other variable definitions. The sample consists of 1,197 firm-quarter observations for which the Twitter tone variables and other variables are available. The *t*-values in parentheses are based on robust standard errors clustered by month of the year. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels, respectively, in two-tailed tests.

Table 9 Predicting Three-day Abnormal Returns around Earnings Announcement Dates: High and Low Information Asymmetry

	Dependent variable = $AR_{(t-1, t+1)}$ around earnings announcement dates					
	Size		Analyst followings		Age	
	Small	Large	Low	High	Young	Mature
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	-0.026	0.033	-0.020	0.010	-0.012	0.000
	(-0.73)	(1.85)	(-0.85)	(0.48)	(-0.66)	(0.02)
Local PosTwt$_{-30,-3}$	-0.765	0.650	-0.669	1.221	-0.678*	0.822
	(-1.28)	(0.78)	(-1.25)	(0.90)	(-1.90)	(0.69)
Local NegTwt$_{-30,-3}$	-0.969*	0.256	-1.137***	0.311	-0.584**	-0.596
	(-2.25)	(0.28)	(-5.87)	(0.65)	(-2.54)	(-1.25)
Nonlocal PosTwt$_{-30,-3}$	0.691	-0.812	-0.040	-0.765	0.782**	-0.793
	(1.84)	(-1.72)	(-0.25)	(-1.52)	(2.32)	(-1.56)
Nonlocal NegTwt$_{-30,-3}$	0.174	-0.269	0.338	-0.322	-0.245	0.240
	(0.97)	(-0.49)	(1.41)	(-0.98)	(-0.81)	(0.98)
Local PosNews$_{-30,-3}$	2.851	-0.124	1.849	0.551	1.077	0.686
	(0.63)	(-0.14)	(0.38)	(0.33)	(0.25)	(0.67)
Local NegNews$_{-30,-3}$	-1.379	0.823	-0.923	0.392	-0.065	0.040
	(-1.67)	(1.61)	(-1.44)	(1.01)	(-0.07)	(0.16)
Nonlocal PosNews$_{-30,-3}$	-4.559	-0.338	-1.601	-1.802	-2.028	-0.977
	(-1.27)	(-0.50)	(-0.46)	(-0.96)	(-1.00)	(-0.55)
Nonlocal NegNews$_{-30,-3}$	1.037	-0.421	0.883	-0.559	0.327	-0.359
	(1.37)	(-1.01)	(0.92)	(-0.89)	(0.57)	(-0.54)
Lagged SUE	-0.020	-0.006	0.007	-0.104*	-0.041	0.002
	(-0.46)	(-0.10)	(0.16)	(-2.01)	(-0.46)	(0.06)
Size	0.006	-0.004	0.005	-0.001	0.003	0.000
	(0.73)	(-1.89)	(1.20)	(-0.65)	(0.95)	(0.06)
BM	0.009	-0.010**	0.007	-0.006	0.002	0.008
	(1.24)	(-2.78)	(0.82)	(-0.93)	(0.29)	(1.02)
Turnover	-0.036	0.004	-0.060	0.022	-0.001	-0.023
	(-0.85)	(0.26)	(-1.39)	(0.73)	(-0.06)	(-0.89)
$AR_{(t-30,t-3)}$	0.016	-0.039	0.029	-0.059**	-0.024	0.061*
	(0.76)	(-0.75)	(1.18)	(-2.92)	(-0.73)	(1.96)
$AR_{(t-2)}$	-0.213	-0.390***	-0.238	-0.232	-0.235	-0.347**
	(-1.37)	(-3.61)	(-1.62)	(-1.54)	(-1.40)	(-2.99)
Adjusted R^2	1.95%	0.06%	0.35%	-0.26%	0.96%	0.16%
N	598	599	601	596	596	601

This table presents the regression estimates of firms' subsequent three-day abnormal returns around quarterly earnings announcement dates on positive and negative words in tweets on Twitter by the extent of information asymmetry. The sample is divided into those with high information asymmetry and those with low information asymmetry based on the sample median of each information asymmetry variable (i.e., firm size, the number of analyst following, and firm age). The dependent variable is the firm's three-day (-1, +1) cumulative abnormal returns centered on the subsequent quarterly earnings announcement date. Abnormal returns are calculated as raw returns adjusted for 25 (5*5) value-weighted portfolios by size and book-to-market (B/M). *PosTwt$_{-30,-3}$* (*NegNews$_{-30,-3}$*) is the average ratio of positive (negative) words to the total number of words between 30 and 3 trading days prior to an earnings announcement date. *Local PosTwt$_{-30,-3}$* (*Local NegTwt$_{-30,-3}$)* and *Nonlocal PosTwt$_{-30,-3}$* (*Nonlocal NegTwt$_{-30,-3}$*) are the average ratio of positive (negative) words to the total number of words in all local and nonlocal tweets about the firm over the same period. *Lagged SUE* is the firm's standardized unexpected quarterly earnings in the previous quarter. It is measured as the seasonal difference in quarterly earnings per share scaled by the end of quarter price. *$AR_{(t-30, t-3)}$* is cumulative abnormal returns from *t*-30 to *t*-3, where *t* is the earnings announcement date. *$AR_{(t-2)}$* is abnormal returns on day *t*-2, where *t* is the earnings announcement date. See Appendix III for other variable definitions. The sample consists of 1,197 firm-quarter observations for which the Twitter tone variables and other variables are available. The *t*-values in parentheses are based on robust standard errors clustered by month of the year. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels, respectively, in two-tailed tests