The Journal of Behavioral Finance & Economics Volume 5, Issues 1&2, 2015-2016, **35-68**

ISSN: 1551-9570

The Effect of Social Media and Gender on the Stock Market

Kyre Lahtinen
Florida State University

Bong Soo Lee
Florida State University

Abstract

Using a unique sample of Twitter posts, also called tweets, we examine the impact of social media on the return, volume, and volatility of the stock market using word list and algorithmic content analysis. We show market returns may be predicted using confidence and sentiment levels. Volume is best predicted by confidence. Volatility is most related to sentiment. We examine one dimension of Twitter user characteristics, namely gender. Our results show that men are more confident and less optimistic than women when they communicate about stocks. We find differences in the ability of communications by men and women to predict market returns, volume, and volatility.

1. Introduction

Social media communication has exploded over the last decade. Use of these networks is not limited to youthful populations. Netflix CEO Reed Hastings was investigated by the Securities and Exchange Commission for possible Reg FD violations for posts he made to a social network in 2012.[1] In 2013 the SEC concluded their investigation by stating that social media is a permissable venue for company announcements provided firms give advanced notice as to which social media outlet will be used to disseminate information.[2] Many firms have taken advantage of the ruling, for example, in October of 2015 Goldman Sachs announced that it would be releasing its quarterly earnings statement through Twitter and not through traditional newswire services.[3] Twitter is amongst the largest social networks in the world. Twitter has more than 200 million active monthly users and they post more than two billion times each month.[4, 5] These posts and user profile data are public, which leads us to ask how this information affects financial markets and whether user characteristics change the informational effects.

This paper addresses the question of the information content of social media. Are posts to social media noise, opinion, or information? We may define noise as communication that has no effects on markets. Noise by definition is, therefore, uninformed and market participants can identify it as such. We may define opinion as the thoughts or preferences of individuals. Opinion, unlike noise, will affect trading behavior. Opinion is the result of the interpretation of available data but reveals nothing new. We may define information as the actual release of facts not previously known to the market. Information will alter the market's expectations about the riskiness or future cash flows of a firm.

We use Twitter as a tool to study the question if social media is noise, opinion, or information, and how social media data affect financial markets. We also examine the effect of one dimension of Twitter users' characteristics, namely gender. We construct a unique sample of over 8 million Twitter posts that include references to 5,500 companies which trade on U.S. stock exchanges over a period from October 1, 2012 to June 30, 2013. We derive a set of language measures from our data using content analysis based on the Loughran and McDonald Financial Sentiment word classification lists. We use these lists to tabulate degrees of sentiment and confidence related communication; additionally, we use a Naive Bayes Classifier to assign buy/sell/hold recommendations to the posts. We aggregate sentiment measures over daily time periods.

We apply dynamic Granger-Causality tests to determine the impact of our language measures on the daily returns and volume of the S&P 500 and the daily closing value of the VIX volatility index. We find confidence and sentiment predict daily returns. Confidence is negatively related to future returns, whereas sentiment is positively related to future returns. Tests of our language measures on daily volume show that only confidence predicts future volume and the relationship is positive. Additionally, we find confidence is positively related and sentiment is negatively related to the expected volatility of the S&P 500, as measured by the daily VIX volatility index.

We find the gender of Twitter users has an effect on the predictive ability of Twitter communication. Again using Granger-Causality tests, we show men's communication more

[1] http://www.businessweek.com/news/2012-12-06/netflix-ceo-hastings-faces-sec-action-over-facebook-post

[2] Report of Investigation Pursuant to Section 21(a) of the Securities Exchange Act of 1934: Netflix, Inc., and Reed Hastings., Release No. 69279 (April 2, 2013).

[3] http://www.wsj.com/articles/goldman-sachs-earnings-are-moving-to-twitter-1444261919?alg=y

[4] http://mashable.com/2012/12/18/twitter-200-million-active-users/

[5] http://www.pcmag.com/article2/0,2817,2364793,00.asp

strongly predicts future returns. Tests of our language measures on the daily volume of the S&P 500 show men's communication more strongly predicts future volume. We find tests of our language measures on the daily level of the VIX volatility index show men's communication more strongly predicts expected future volatility. Tests of the differences in mean language measures suggest that men are more confident and have less positive sentiment than women in their investment communications.

Increasingly there is more interest in finance and economics on the role internet information and gender play in investor and consumer behavior. We add to this literature by connecting stock specific language appearing on social media to the market. We also add to the literature by examining the communication that men and women have about the stock market and how that communication might differ between genders.

This article is organized in the following manner. Section **Error! Reference source not found.** reviews related literature. Section **Error! Reference source not found.** outlines the data used in this study and its sources. Section 4 contains an empirical analysis of the questions at hand. Section

Table XIII: Tests of Differences in Sentiment

This table contains results for t-tests and Wilcoxon-Mann-Whitney rank-sum z-scores for differences in our language measure. Differences are calculated as the level for women minus the level for men. Bull is bullishness. Conf and Sent are the DataSift confidence and sentiment measures. Strg, Weak, Unct, Neg, and Pos are scores for strong, weak, uncertainty, negative, and positive language from the Loughran and McDonald dictionary. SentLM is the sentiment score based off of the Loughran and McDonald Dictionaries.

Var	Difference	T-Stat	P-Value	Z-Score	P-Value
Panel A - Confidence					
Conf	-1.4024	-7.8698	0.0000	-76.659	0.0000
Strg	-0.0751	-15.0390	0.0000	-44.669	0.0000
Weak	-0.0333	-5.2439	0.0000	-15.718	0.0000
Unct	-0.0788	-6.5466	0.0000	-25.243	0.0000
Panel B - Sentiment					
Bull	0.1270	16.0200	0.0000	8.3308	0.0000
Sent	0.0209	0.5647	0.5730	0.7676	0.4427
SentLM	-0.0832	-1.8622	0.0632	-0.3327	0.7394
Neg	0.2450	6.3812	0.0000	3.1148	0.0018
Pos	0.1614	5.0725	0.0000	4.9846	0.0000
Liti	0.1140	8.9540	0.0000	5.9913	0.0000

5 concludes the paper.

2. Literature Review

Investor sentiment, communication, and information are known to be important to financial markets. Baker and Wurgler (2007) argue that one way to measure investor sentiment is by gauging investor mood. Twitter data represents a direct measure of investor opinion and mood, much like the Google Search Volume Index which is used by Da, Engelberg, and Gao (2011) as a direct measure of investor attention. Others have used internet based user generated data to study finance research questions. Das and Chen (2007) use internet message board posts to extract small investor sentiment. Zhang and Swanson (2008) use internet message boards to examine bias in day traders. Mizrach and Weerts (2009) use internet chat rooms when studying investor skill. And, Sabherwal, Sarkar, and Zhang (2011) use message boards to study the ability of users to manipulate certain stocks.

We anticipate that confidence will be negatively related to returns and positively related to trading volume. This is based on the work of Barber and Odean (2001), Statman, Thorley, and Vorkink (2006), and Asem and Tian (2010). These studies use monthly data whereas we use daily data. We use daily data due to the disposition and length of our data set. We are comfortable with this decision given that Chou and Wang (2011) examine overconfidence using daily data.

We expect gender to influence the predictive ability of communication based on psychological evidence. An exhaustive study by Costa, Terracciano, and McCrae (2001) examines differences between personality traits in men and women. The authors utilize the Revised NEO Personality Inventory of Costa and McCrae (1992). This is a survey based assessment of 30 personality traits across five orthogonal dimensions, namely neuroticism, extraversion, agreeableness, openness to experience, and conscientiousness. Costa, Terracciano, and McCrae (2001) aggregate data from studies of men and women of 26 different cultures. They find that men are more assertive and more confident than women. They also show that countries with high levels of individualism as ranked by Hofstede (1998), like the United States, have greater differences in gender attributes. Based on this evidence, we expect men will be more confident than women in their communication about stocks.

In addition, Costa, Terracciano, and McCrae (2001) show that women are more susceptible to neurotic personality traits than men. This means that women are more likely to suffer from conditions like depression, anxiety disorder, and panic disorder. The findings of Lin and Raghubir (2005) offer complementary results by showing that men are more optimistic than women and are less likely to change their beliefs when given new information. We expect to find women communicate more negatively about stocks.

Finance and economics literature reinforces our expectations about gender differences and the role these differences may play in their communication and the impact of that communication on markets. Agnew (2006) shows women are less subject to behavioral biases in retirement savings. Agnew, Balduzzi, and Sunden (2003) and Halko, Kaustia, and Alanko (2012) show men are less risk averse than women. Barsky, Juster, Kimball, and Shapiro (1997) find men have a higher tolerance for risk. Sunden and Surette (1998) show women are less likely to hold equities than men. Säve-Söderbergh (2012) find when engaging in more risky behavior, men take more risk than women. However, Schubert, Brown, Gysler, and Brachinger (1999) contradict these results and find that women do not make less risky choices than men. We expect men's communication to be more related to future volatility given men's propensity to be more risk seeking.

Lusardi and Mitchell (2008) show women have a lower level of financial literacy relative to men. The lower level of literacy is affected by educational choices men and women make and differences in self image. Steele, James, and Barnett (2002) report that women in men dominated university majors are more likely to change their major. Their results also suggest that there are obstacles to women in math, science, and engineering related fields. Nosek, Banaji, and Greenwald

(2002) examine how men and women identify with different roles and the impact that identification can have on life choices. Men are shown to identify more with math while women are shown to identify with the self. Even women who select quantitative intensive fields have a difficult time identifying math as part of their identity. Schmader (2002) shows that women who place a very high importance on their gender identity are more disposed to perform worse on math exams.[6] These facts build a case that leads us to expect men to communicate more frequently than women about stocks.

Barber and Odean (2001) model investor behavior based on evidence that men are more overconfident than women. They find that men are more likely to trade excessively relative to women and that men's stock returns are more likely to be negatively affected by overconfidence and excessive trading. They are motivated by psychological research, showing that differences in overconfidence between men and women are task dependent (Lundeberg and Fox, 1994), and, therefore, men are more likely to be overconfident in certain situations. Men are known to be more overconfident than women in the financial setting (Lewellen, Lease, and Schlarbaum, 1977). Generally, we anticipate that confidence levels will be negatively related to returns and positively related to trading volume and the significance of these relationships is likely more pronounced in men.

We extend prior literature by examining whether our language measures are dynamically related to the stock market. Our tests differ from prior literature in several important ways. Our language measures use Twitter data that is only related to tweets about NYSE, Amex, and Nasdaq firms. Past literature has data focused on any tweet with words that directly describe peoples' feelings. Our sample is more likely to capture sentiment that relates to trading behavior. We examine confidence and sentiment related language using finance specific dictionaries; previous studies have used more general content analysis techinques. As we are looking at only stock related tweets, our dictionaries are more likely to accurately classify stock market specific language. We believe that performing out of sample tests on online and social media are important for confirming the role of online and social media's effect on the market.

We extend prior work by performing a more thorough analysis of the relationship between online content and the stock market. We use the S & P 500 as our market proxy making our proxy more broad based and more representative of the entire market than market proxies used in some previous work. In addition to analyzing returns, we also study the dynamic causal relationship between tweets and volume and volatility. This represents the first time this has been examined using social media. We identify differences across user characteristics, the way men and women communicate differently, and the predictive ability of each gender. We are able to demonstrate whether language used on Twitter conforms to the relationships and expectations of prior literature. We believe this is a key feature to understanding the mechanism that lies beneath the influence of Twitter on the market.

3. Data

We use social media data as a tool to examine the relationship between investor communication and financial markets. Social networking data and the data generated by its users have been used

[6] Being curious about differences in grade performance in finance courses, we examine the differences in gender grade performance by acquiring a dataset of grades for two introductory finance courses at a major public university. The dataset covers a financial management course and a financial markets course for semesters between 2008 and 2013. Our tests conclude that men outperform women in these courses at a statistically significant level.

in prior gender studies (Thelwall, Wilkinson, and Uppal, 2009) and in finance related work (Bollen, Mao, and Zeng, 2011). Antweiler and Frank (2004) use the information content of internet message board posts to explain stock market returns, volume, and volatility. It is our belief that using social media data, in other words, the actual communications that occur between individuals, we will be better able to gain new insight into stock transactions. While there are some studies, including those listed in the literature review, that utilize data on the realized trades of individuals, there is no way to confirm the attitude of that person towards the stock or trade in question. We believe that social media data can provide new insights not previously explored by examining aggregate financial communication and the market. We have chosen Twitter as the most appropriate social networking platform. In fact, Bollen, Mao, and Zeng (2011) use Twitter data to predict stock market movements. The data used by Bollen, Mao, and Zeng (2011) differ from ours in a meaningful way to be discussed later. An explanation of what Twitter is, how it functions, and why it is most suited for our needs follows.

3.1. Twitter

Twitter is a social media company specializing in microblogging.[7] Twitter posts, also called tweets, can be no more than 140 characters. Twitter has more than 200 million active monthly users and they tweet more than two billion times each month. Twitter is unique in nature among blogging and microblogging platforms in that the tweets are meant to be consumed as they happen in quasi-real time, rather than be saved and consumed later in the way that most news or long form blogging is meant to be consumed.[8] Figure I is an example of one such tweet.

Figure I

To facilitate short-form communication, Twitter users have created several conventions for relaying specific types of information in tweets; the three most useful follow next. The first defines if a tweet is directed at a specific user. If this is the case, the tweet will contain an @ mention (pronounced "at" mention). This takes the form of the @ symbol followed by the username of the person. For example, including "@federalreserve" would direct a user's comments to the Federal Reserve's Twitter account. The second convention identifies the main subject, topic, or a related issue to a tweet. This is accomplished with the use of the # (pronounced "hash tag") symbol. This takes the form of the # symbol followed by the topic or keyword of the tweet. Including "#inflation" in a tweet lets others know the focus of your tweet. The third convention specifies if the information in the tweet is related to a particular stock. The tweet will contain a $ followed by the ticker symbol of the company in question. Putting "$JPM" identifies the tweet as connected to the user's opinion of the stock of JP Morgan Chase and Co. This convention is particularly useful for financial research because it isolates tweets that are related to an individual's financial opinion. A variety of

[7] Microblogging is a form of online publication that is subject to significant length and/or size constraints.

[8] Twitter does not natively provide a mechanism for its users to see which tweets they have and have not read. It is assumed in the tech community that Twitter wants its users to read as things happen rather than catching-up at some later time as is customary with e-mail.

other information is also generated when a user tweets. This includes tweet content, user location, and other metadata about the user.

Bollen, Mao, and Zeng (2011) set the precedent for using Twitter data in financial research. Their work has had significant influence since its publication. This article has been cited more than 300 times in newspapers, magazines, and academic research since its original publication. It is directly responsible for the formation of a hedge fund that was run by Derwent Capital.[9] Commercial providers of Twitter data advertise its usefulness in creating trading and other financial strategies.[10], [11]

If there was any doubt about the impact that Twitter and tweets can have on the market, observe the market reaction surrounding a single falsified tweet from the official Twitter account of the Associated Press on April 23rd, 2012 in Figure II below. Their account was hacked and nefarious persons tweeted that The White House had experienced an explosion and President Obama had been injured. The stock market reacted nearly instantaneously to this news. The S&P 500 fell 0.924% between 1:08:10 PM and 1:10:05 PM. The Associated Press quickly announced that the tweet was a fake and the market recovered to its earlier level by gaining 0.896% between 1:10:05 PM and 1:12:55 PM. The markets are paying attention to what is on Twitter. We, therefore, conclude that there is enough precedent to use data of this kind for our research.[12]

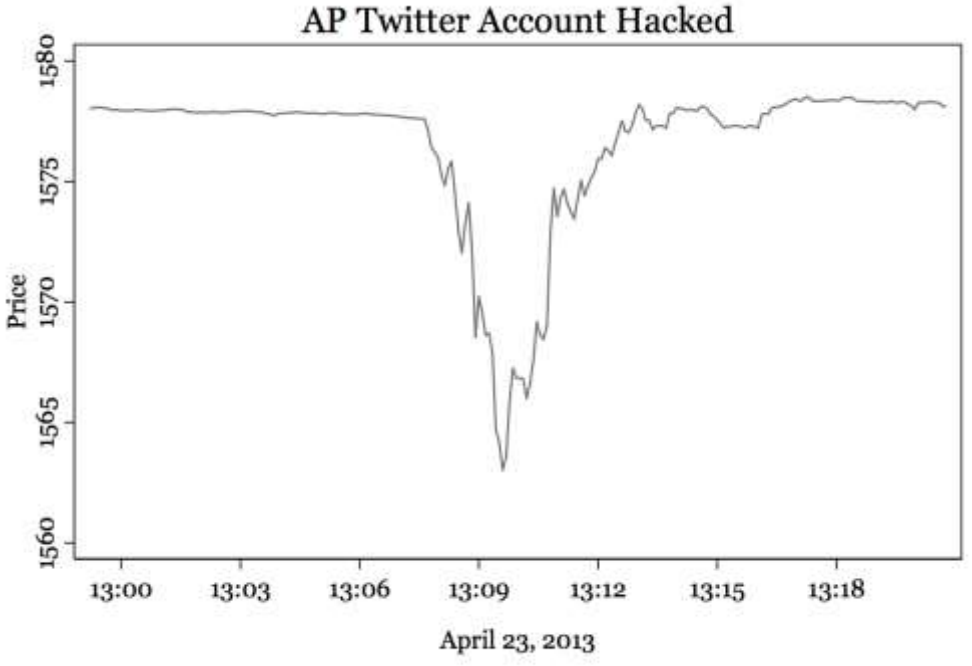

Figure II

[9] http://www.theatlanticwire.com/business/2011/08/how-twitter-based-hedge-fund-beat-stock-market/ 41389/

[10] http://gnip.com/product_overview/

[11] http://datasift.com/solutions/industries

[12] It is possible that the nearly instantaneous effect of this episode on the market was driven by institutional trading. Institutions have the resources to monitor, either by individuals or by algorithm, the activity on social media. Institutions are also more likely to engage in high frequency trading. The possibility of institutions being the driving cause behind our results does not alter our main conclusions.

3.2. Data Description

Twitter data may be collected from several sources.[13] It is available directly from Twitter through use of application programming interfaces (API). The free API for collecting tweets is called the "garden hose." This grants access to a random sample of approximately 10% of the real-time stream of tweets. The complexity and size of data filters used on the garden hose are limited. The API that allows access to the full stream of user generated tweets is called the "fire hose." Access to the fire hose is extremely limited. Use of either of these APIs collects data in real-time. It is not possible to query historical tweets. Historical Twitter data is commercially available through firms such as DataSift. DataSift is a provider of aggregated social media data and analytical services.

3.3. Sample Data

We have acquired a unique dataset of tweets from DataSift for a period from October 1, 2012 thru June 30, 2013. This data includes tweets from the entire Twitter fire hose (all publicly available tweets) filtered for those tweets that contain references to the top 5,500 companies listed on the NYSE, Nasdaq, and Amex.[14] The filter is built around references to a company's name or the $ convention discussed above. DataSift provides several additional features not contained in the raw Twitter data. DataSift analyzes tweets to determine the gender of the user, and analyzes tweets for sentiment and confidence levels. We do not exclude tweets that contain a

Table I: Summary Statistics

[13] The Library of Congress has received the entire Twitter fire hose since its inception (http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/), but it has had significant difficulty formatting the data in such a way to make it available to interested parties (http://www.digitaltrends.com/social-media/library-of-congress-useless-twitter-archive-is-almost-complete/). Visiting the Library of Congress website shows that there is an intense interest by academics in utilizing Twitter data.

[14] Gonzalez-Bailon, Wang, Rivero, Borge-Holthoefer, and Moreno (2012) contend that using the search function of Twitter overrepresents central users and therefore biases data. We feel that it is important to access the full fire hose of Twitter data in order to construct the best data. Research that does not originate from the fire hose should be careful to guard against such possible biases.

This table contains summary statistics on tweets collected from October 1, 2012 until March 31, 2013. Tweets represents the total number of tweets. Gender is a binary variable that is set to 1 if the user's name is determined to be at least somewhat male and 0 for at least somewhat female. Conf measures the confidence of the language used in the tweet as provided by DataSift. Strg is the score of words that fall into the Loughran and McDonald strong modal words group. Weak is the score of words that fall into the Loughran and McDonald weak modal words group. Unct is the score of words that fall into the Loughran and McDonald uncertainty words group. Sent measures the positive or negative sentiment of the tweet as provided by DataSift. SentLM represents the difference between positive and negative word score using the Loughran and McDonald positive and negative word groups. Neg is the score of words that fall into the Loughran and McDonald negative words group. Pos is the score of words that fall into the Loughran and McDonald positive words group. Liti is the score of words that fall into the Loughran and McDonald litigious words group. The data was collected using DataSift filtering all tweets for the NYSE, Amex, and Nasdaq firms that are referenced by their name or ticker using the dollar sign convention. Ret is the daily return on the S&P 500. Vol is the daily volume of the S&P 500 in millions of shares. Vix is the daily closing value of the VIX volatility index.

Variable	N	Mean	Median	Min	Max	StDev
Tweets	8342229	-	-	-	-	-
Gender	3818022	0.7655	1.0000	0	1.0000	0.4237
Conf	8342221	84.6533	100.0	26.0	100.0	18.4763
Strg	8342229	0.2638	0.0000	0.0000	44.6300	1.7708
Weak	8342229	0.3002	0.0000	0.0000	48.2151	1.9823
Unct	8342229	0.6230	0.0000	0.0000	66.9200	3.0800
Sent	2417167	1.1525	3.0000	-22.0000	35.0000	4.6995
SentLM	8342229	-0.9163	0.0000	-108.6239	81.1933	7.8125
Neg	8342229	2.2027	0	0.0000	108.6239	6.4688
Pos	8342229	1.2864	0.0000	0.0000	81.1933	4.3790
Liti	8342229	0.4451	0.0000	0.0000	92.6614	2.9408
Ret	186	0.0006158	0.0007815	-0.02501	0.02540	0.007841
Vol	186	3500	3490	1250	5800	540
Vix	186	15.0478	14.68	11.3	21.79	2.1007

hyperlink. Bollen, Mao, and Zeng (2011) exclude tweets with hyperlinks as a way to exclude spam messages. Results from Hanke and Hauser (2008) suggest spam messages have an impact on financial markets and provide a basis for leaving spam messages in the sample. Of the tweets in our sample 71.6% contain at least one hyperlink. In the process of manually categorizing tweets, we discovered that a large portion of tweets that contain hyperlinks are not spam. Also, unlike Bollen, Mao, and Zeng (2011), we do not filter out retweeted messages.[15] We retain retweets because they still reflect the opinion of the users who retweeted it. A summary of this data can be found in Table I.

Table II contains the 40 most commonly mentioned firms. Percent is calculated as the number of times a firm appears in the sample divided by the total number of firm references. Count is the

[15] Retweets are instances where a user repeats the content of another user's tweet. Retweeted messages include the name of the originator of the content.

raw count of references to a firm. In total there are 5,846 unique firms mentioned across 8,342,229 tweets. Highly mentioned firms include large, small, and well-known firms.

Table II: Firm Frequencies

This table contains the most commonly mentioned firms. Percent is calculated as the number of times a firm appears in the sample divided by the total number of firm references. Count is the raw count of references to a firm.

Company Name	Percent	Count	Company Name	Percent	Count
Amazon.com, Inc.	0.0079	53483	Intel Corporation	0.0041	28233
Apple Inc.	0.0817	556751	J P Morgan Chase & Co	0.0042	28518
ArcelorMittal	0.0037	24928	Life Time Fitness	0.0036	24483
Bank of America Corporation	0.0061	41817	LinkedIn Corporation	0.0026	17748
Boeing Company	0.0051	34461	McDonald's Corporation	0.0026	18031
Citigroup Inc.	0.0034	23397	Microsoft Corporation	0.0116	78994
Colgate-Palmolive Company	0.0031	21351	Morgan Stanley	0.0192	130444
ConocoPhillips	0.0036	24557	Netflix, Inc.	0.007	47649
Dell Inc.	0.0054	36864	News Corporation	0.0134	91206
eBay Inc.	0.0029	19460	Nokia Corporation	0.0026	17630
EnergySolutions Inc	0.0028	19051	Oracle Corporation	0.003	20446
Facebook, Inc.	0.0201	136734	PowerShares QQQ Trust, Series 1	0.0042	28731
Ford Motor Company	0.0062	42370	Research in Motion Limited	0.007	47631
General Motors Company	0.0029	19876	Sears Holdings Corporation	0.0029	20032
Goldman Sachs Group, Inc.	0.0061	41313	Starbucks Corporation	0.0029	19609
Google Inc.	0.0218	148646	Tiffany & Co.	0.0029	19804
Greenhill	0.0027	18665	Vringo, Inc.	0.0033	22135
Groupon, Inc.	0.003	20255	Walt Disney Company	0.0078	53457
Herbalife LTD.	0.0044	30159	Yahoo! Inc.	0.0037	25256
Hewlett-Packard Company	0.0028	19152	Zynga Inc.	0.0046	31020

3.4. Content Analysis

3.4.1 Classification with Dictionaries

We analyze the tweets using a "bag of words" methodology as described in Loughran and McDonald (2011). Under this process each tweet is divided into word occurrence vectors with the unique words appearing in each tweet and the count of each unique word. Next the words are tagged against a dictionary. Finally, the words are given weights to determine the strength of any given measure within the text.

For this study we use the Loughran and McDonald Financial Sentiment word classification lists. Prior to Loughran and McDonald (2011), many studies used the Harvard-IV-4 TagNeg (H4N) file from the more broadly known Harvard Psychosociological Dictionary.[16] Loughran and McDonald (2011) determine that the more general Harvard dictionary is not appropriate for analyzing financial texts. The Loughran and McDonald dictionary was created by extracting words

[16] Available at http://www.wjh.harvard.edu/~inquirer/

that appear in 10-Ks and assigning them to categories of positive, negative, litigious, uncertainty, strong, and weak words. A finance specific list more correctly interprets a word like vice as identifying a position within a company, such as vice-president, rather than identifying it as a negative and possibly addictive behavior, such as alcoholism.

Negative and positive words are verified for their proximity to negations, words like "not". A positive word that has a negation within the three previous words is counted as negative. A negative word with a negation in the prior word is counted as positive.

Weighting systems for bag of words analysis account for several different aspects of a single document and of the corpus of documents, where a tweet is the document and the sample of tweets is the corpus. We use the foundations of the weighting schemes found in Chisholm and Kolda (1999) for our weights. Let W_{ij} be the weight of term *i* in document *j*. The weight is comprised of three terms as indicated in equation (1). L_{ij} is the local weight for term *i* in document *j*. We make L_{ij} equal to the simple frequency of each word in a document. This approach is appropriate when there is low probability of a word occurring more than twice within a document as we observe in our corpus. G_i is the global weight for term *i*, where *N* is the number of words in the corpus and n_i is the frequency of word *i* in the corpus. This term adjusts the weight of a word for impact. The value for G_i is inversely related to its usage; the more a word is used the lower the impact. N_j is the normalization factor for document *j*; this is used to compensate for differences in document length. We set N_j equal to 1 as a simplifying assumption. We justify this choice because tweets have a hard upper limit on length, 140 characters; therefore, by construction all tweets are some type of short form communication. We recognize that this may be viewed as a strong assumption. Alternative specifications for document length, N_j, do not meaningfully alter our results.

$W_{ij}=L_{ij}G_iN_j$,

where:

$$L_{ij}=f_{ij}\ , \quad G_i=log\left(\frac{N}{n_i}\right), \text{ and } \quad N_j=1 \quad . \tag{1}$$

The score for document *k* for word list *l* then becomes the summation of the weights for each individual word *i* within .

$$Score_{kl}=\sum_{i=1}^{n} W_{il}. \tag{2}$$

3.4.2 Naive Bayes Classifiers

There is some concern that the language that appears on Twitter may not conform to standard dictionaries, whether it be the Harvard, or Loughran and McDonald dictionary. Writing on Twitter, like texting on a phone, has strict limitations on message length. Varnhagen, McFall, Pugh, Routledge, Sumida-MacDonald, and Kwong (2009) describe new language and spelling in instant messages. On Twitter in addition to the use of common online and text short hand, like "lol" for "laugh out loud," other altered spellings have occurred, like "bot" for "bought." To more readily account for new online language we use natural language analysis tools to classify the text. Natural

language analysis uses algorithms to identify the most important components of a word, sentence, or passage that link it to a predefined category.

We use the Naive Bayes algorithm as in Antweiler and Frank (2004). This algorithm may also be referred to as simple Bayes, independence Bayes, or idiot's Bayes. The algorithm is said to be naive because it assumes the occurrence of each word is independent. This assumption is obviously strong and unlikely to hold true in actual communication. Nevertheless, the algorithm performs well.

We use the Natural Language Toolkit for the Python programming language to process our text.[17] The process of classifying text follows a four step process. First, manually classify a subset of documents. The manually classified subset is split into two parts, a training set and a test set. Second, the training data set is used by the algorithm to create a set of rules that identify each manually specified category. Third, the rules are then applied to the test set to see how accurate the algorithm is at classifying the documents. The training set should not be used to verify accuracy of the classifier. The rules may be adjusted until a satisfactory level of accuracy is achieved. Finally, the rules are applied to the entire sample of documents.

Wu et al. (2008) show Naive Bayes classifiers can be represented rather simply in general probabilistic terms. We demonstrate the intuition of the classifier using a two classification model for simplicity. Let the type or classification of a document be $T=\{A,B\}$. Let a vector of words from the document, in our case tweet, be $W = \{w_1, w_2, \ldots, w_{n-1}, w_n\}$. Therefore, $P(T|W)$ becomes the probability a given vector W is of type T. To properly assign a type to a document any monotonic function of $P(T|W)$ is appropriate. We follow Wu et al. (2008) and use equation (3) to demonstrate one such appropriate equation. Let $f\left(W|T_i\right)$ be the marginal distribution.

$$\frac{P\left(A|W\right)}{P\left(B|W\right)}=\frac{f\left(W|A\right)P\left(A\right)}{f\left(W|B\right)P\left(B\right)} \quad . \tag{3}$$

As shown in Wu et al. (2008), the marginal distributions, $f\left(W|T_i\right)$, of our data are discrete. In each document you then have the ability to simply count the proportion of each word of type T_i. Since each word is independent, equation (3) can be re-written as:

$$\frac{P\left(A|W\right)}{P\left(B|W\right)}=\frac{\prod_{j=1}^{n} f\left(w_j|A\right)P\left(A\right)}{\prod_{j=1}^{n} f\left(w_j|B\right)P\left(B\right)}=\frac{P\left(A\right)}{P\left(B\right)}\prod_{j=1}^{n}\frac{f\left(x_j|A\right)}{f\left(x_j|B\right)} \quad . \tag{4}$$

Equation (4) can be more easily read as a summation, therefore, we take the log transformation giving us the straightforward result in equation (5).

$$\ln\frac{P\left(A|W\right)}{P\left(B|W\right)}=\ln\frac{P\left(A\right)}{P\left(B\right)}+\sum_{j=1}^{n}\ln\frac{f\left(x_j|A\right)}{f\left(x_j|B\right)} \quad . \tag{5}$$

We refer the interested reader to Wu et al. (2008), Friedman, Geiger, and Goldszmidt (1997), Bird, Klein, and Loper (2009), and Antweiler and Frank (2004) for more thorough discussion on the Naive Bayes classifier, alternative algorithms to the classifier, and possible modifications to the classifier.

[17] Available at http://nltk.org/

Using the classified text, we construct a measure of bullishness to capture the overall attitude of Twitter users toward the direction of the stock market. In equation (6), let *t* represent daily time intervals and N^i be the number of tweets that are rated either buy, sell, or hold. The variable $Bull_t$ may take values over the interval [-1,1], where positive (negative) values indicate bullish (bearish) sentiment. We elect to include the number of hold recommendations as a form of dilution to bullish or bearish sentiment.

$$Bull_t = \frac{N^{Buy} - N^{Sell}}{N^{Buy} + N^{Hold} + N^{Sell}} \quad . \tag{6}$$

We manually classify 10,000 tweets randomly selected from our sample. Our training set uses 9,000 tweets and the test set uses the remaining 1,000.[18] we rate these tweets as indicating "buy", "sell", "hold", or "no" action implied. Tweets rated as "buy" include those where the authors specifically indicate that they purchased or plan to purchase shares; or, the author expresses an opinion that the stock price will rise. Tweets rated as "sell" include those where the authors specifically indicate that they sold or plan to sell shares; or, the author expresses an opinion that the stock price will fall. Tweets rated as "hold" include those where the authors indicate that they would not change their current position; or, the author expresses an opinion that the price of the stock will not change. Tweets rated "no action implied" are those where no specific opinions about changing stock price are contained in the tweet. For example, a tweet may contain positive information about a company, but if the author does not specifically state how that information will change the price of the stock, it is rated "no".

Table III: Buy/Sell/Hold

This table contains information for Naive-Bayes classification of tweets. The column 1 contains our manual classification of tweets used in our training data set. Column 2 contains the manual rating for tweets used in our test data set. Column 3 contains the guessed classification using the classifier. Column 4 contains the rating for all tweets in the sample.

	(1) Rated Train	(2) Rated Test	(3) Guessed Test	(4) Guessed Sample
Buy	721	92	44	380868
Sell	299	55	42	404944
Hold	90	8	13	104803
Non	7890	845	901	7451614
Total	9000	1000	1000	8342229

3.4.3. DataSift

DataSift provides content analysis services. As part of these services they analyze tweets to determine the gender of the user based upon a database they maintain of male and female names.

[18] Splitting the manually classified data into 90% training set and 10% test set is recommended by Bird, Klein, and Loper (2009).

DataSift has also provided sentiment and confidence analysis of each tweet. These scores are calculated using DataSift's proprietary method.

3.5. Variables

All variables represent aggregated measures unless otherwise specified. We aggregate our language measures by taking daily averages of the indicated measure over each day, where day is defined as d_t. The time interval for d_t is the period from the close of the market on d_{t-1} until the close of the market on d_t. If d_{t-1} is a non-trading day, the interval goes back to the close of the market on the last trading day. This ensures that our language measures match the same interval as the market related measures. This also prevents the loss of observations in our Granger-Causality tests due to gaps in the time series.

Ret and *Volume* are the daily return and volume, respectively, on the GSPC S&P 500 Index with data collected from Yahoo! Finance. *Vix* is the daily closing value of the VIX Volatility Index with data collected from Bloomberg.

Conf measures the confidence of the language used in the tweet as provided by DataSift. *Strg* is the measure of words that fall into the Loughran and McDonald strong modal words group. *Weak* is the measure of words that fall into the Loughran and McDonald weak modal words group. *Unct* is the measure of words that fall into the Loughran and McDonald uncertainty words group.

Sent measures the positive or negative sentiment of the tweet as provided by DataSift. Positive values represent positive sentiment. A value of 0 represents neutral sentiment. Negative values represent negative sentiment. *SentLM* represents the difference between positive and negative word measures using the Loughran and McDonald positive and negative word groups. *Neg* is the measure of words that fall into the Loughran and McDonald negative words group. *Pos* is the measure of words that fall into the Loughran and McDonald positive words group.

Gender is a binary variable that is set to 1 if the user's name is determined to be at least somewhat male and 0 for at least somewhat female. Name classifications are provided by DataSift. *Liti* is the measure of words that fall into the Loughran and McDonald litigious words group. *Bull* is the measure of bullishness described in a prior section.

It is important to note that our measures of sentiment, like many that have been used previously in finance literature, do not attempt to identify the various types of sentiment such as Affect, Mood, or Emotions. These types of sentiment vary in cause, duration, intensity, and function. This presents an opportunity for future research to further define and understand the impact sentiment has on investing.

4. Empirical Results

4.1. Predicting Return, Volume, and Volatility

We test the dynamic relations between lagged language measures and daily returns, volume, and volatility of the S&P 500 using Granger Causality tests. Equation (7) shows the model we estimate. Variable X represents our sentiment measures, and our dependent variable Y represents *Ret*, *Vol*, and *Vix*. We report coefficients and p-values for the F-statistic associated with equation (7).

$$Y_t = \alpha_0 + \sum_{i=1}^{5} \alpha_i Y_{t-i} + \sum_{j=1}^{5} \beta_j X_{t-j} + \varepsilon_t \ . \quad (7)$$

Variable X_t is said to Granger-cause variable Y_t if we reject:

$$H_0: \beta_j = 0, \text{ for all } j \text{ in (7).} \quad (8.1)$$

In other words, X_t Granger-causes Y_t if lagged X_t can predict current Y_t , controlling for past Y_t. If the null H_0:β_j=0, for all j in (7) is rejected, we find evidence that X_t Granger-causes Y_t . This specification also allows us to test the net (cumulative) effect of lagged X_t on Y_t. Specifically, we test:

$$H_0: \sum_{j=1}^{5} \beta_j = 0 \text{ in (7),} \quad (8.2)$$

which allows us to test for the sign of the causal relation. If we find the net effect (sum of coefficients, $\sum_{j=1}^{5} \beta_j$) is significantly positive, the sign of the causation is positive.

For ease in interpreting results, we perform the standard normal transformation on all variables such that they all have mean equal to zero and unit variance. Therefore, a coefficient on *X* of 1 signifies that for every one standard deviation change in *X* there is a one standard deviation change in the value of *Y*. For example, the standard deviation of *Ret* from Table I is 0.7841%.

Table IV contains our results for causality tests using all tweets where the dependent variable is *Ret*. Five of our ten language measures, across confidence and sentiment, are significant. *Conf* is significant and negatively related to future returns. This is consistent with the idea that future returns are negatively related to confidence. The total effect of lagged *Conf* is -0.2876; a one standard deviation change in *Conf* leads to -0.2876 standard deviation change in *Ret*. *Weak* is significantly related to future returns. Its net effect is positive. This is consistent with the results for *Conf* because greater levels of weak words decrease an individual's confidence. It appears that low confidence is driving the results for *Conf* given *Strg* is insignificant. The magnitude for the aggregate effect of *Weak* is 0.1196. This amount is economically meaningful.

Table IV: Predicting Return Using All Tweets

This table contains results for Granger-Causality tests where the language measures use all tweets in the sample. All variables have been standardized with mean zero and unit variance. Dependent variable, *Y*, is the daily return of the S&P 500. Coefficients on *Y* are denoted as α_{t-i}. Independent variables, *X*, are our language measures. Coefficients on *X* are denoted as β_{t-i}. Bull is bullishness. Conf and Sent are the DataSift confidence and sentiment measures. Strg, Weak, Unct, Neg, and Pos are scores for strong, weak, uncertainty, negative, and positive language from the Loughran and McDonald dictionary. SentLM is a sentiment score based off of the difference between positive and negative word usage from the Loughran and McDonald Dictionary.

Coefficient	(1) Bull	(2) Conf	(3) Strg	(4) Weak	(5) Unct	(6) Sent	(7) SentLM	(8) Neg	(9) Pos	(10) Liti
α_{t-1}	-0.0752	-0.0394	-0.0938	-0.111	-0.0918	-0.0836	-0.0866	-0.0912	-0.0836	-0.0938
	(0.324)	(0.602)	(0.218)	(0.139)	(0.226)	(0.275)	(0.255)	(0.231)	(0.272)	(0.216)

α_{t-2}	-0.0584	-0.0767	-0.0743	-0.0612	-0.0418	-0.0483	-0.0190	-0.0210	-0.0521	-0.0916
	(0.442)	(0.306)	(0.322)	(0.410)	(0.578)	(0.526)	(0.800)	(0.779)	(0.491)	(0.225)
α_{t-3}	-0.00684	-0.0696	-0.0253	-0.00443	-0.00403	-0.0166	-0.0284	-0.0467	-0.0157	-0.0283
	(0.929)	(0.361)	(0.737)	(0.953)	(0.957)	(0.825)	(0.700)	(0.523)	(0.836)	(0.704)
α_{t-4}	-0.118	-0.0779	-0.125*	-0.127*	-0.126*	-0.119	-0.110	-0.118	-0.126*	-0.137*
	(0.124)	(0.311)	(0.097)	(0.090)	(0.097)	(0.113)	(0.134)	(0.105)	(0.099)	(0.069)
α_{t-5}	-0.154**	-0.181**	-0.160**	-0.181**	-0.179**	-0.169**	-0.160**	-0.170**	-0.162**	-0.168**
	(0.048)	(0.015)	(0.037)	(0.017)	(0.020)	(0.027)	(0.032)	(0.021)	(0.034)	(0.028)
β_{t-1}	0.0299	-0.239***	-0.00322	0.185**	0.0590	0.0268	0.0342	-0.102	-0.109	-0.0878
	(0.702)	(0.001)	(0.968)	(0.034)	(0.445)	(0.734)	(0.654)	(0.171)	(0.154)	(0.239)
β_{t-2}	0.00424	0.0671	0.0649	0.157*	0.166**	-0.127	-0.168**	0.203**	0.0875	0.00279
	(0.958)	(0.399)	(0.423)	(0.076)	(0.036)	(0.126)	(0.032)	(0.010)	(0.259)	(0.971)
β_{t-3}	-0.0395	-0.0537	0.0239	0.0343	0.00301	0.149*	0.183**	-0.204**	-0.0293	-0.187**
	(0.628)	(0.506)	(0.765)	(0.703)	(0.970)	(0.078)	(0.022)	(0.011)	(0.707)	(0.015)
β_{t-4}	0.0810	-0.173**	0.158*	-0.0857	-0.0899	0.106	0.179**	-0.178**	-0.0656	-0.0329
	(0.317)	(0.040)	(0.052)	(0.331)	(0.257)	(0.222)	(0.027)	(0.029)	(0.398)	(0.670)
β_{t-5}	-0.0656	0.111	0.0384	-0.171*	-0.0754	0.0569	-0.0195	-0.0194	-0.0732	-0.0839
	(0.405)	(0.189)	(0.638)	(0.053)	(0.340)	(0.511)	(0.808)	(0.809)	(0.350)	(0.272)
Sum of β_{t-i}	0.01004	-0.2876	0.28198	0.1196	0.0627	0.2117	0.2087	-0.3004	-0.1896	-0.3888
Prob > F	0.8984	0.0033	0.2953	0.0233	0.1686	0.1701	0.0078	0.0009	0.3855	0.0989
N	181	181	181	181	181	181	181	181	181	181

Of our sentiment measures, *SentLM* predicts future returns. The net effect of *SentLM* is 0.2087. Again, this amount is economically large as well as statistically significant. Insight can be gained into the driving force behind these results from our measures for positive and negative word usage. We see that negative words, as measured by *Neg* and *Liti*, have predictive value of future returns whereas positive words do not predict returns. *Neg* and *Liti* are both statistically and economically significant with their net effects being negative.

Table V has results where *Vol* is the dependent variable. Aggregate tweets have less ability predicting future volume than they do predicting future returns. There are three confidence

language measures that are significantly related to *Vol*. *Conf* is significant with the net effect being 0.3695; as people are more confident they increase their trading activity. *Weak* and *Unct* are also significant with a net effect of -0.047 and -0.0798, respectively. As each of these measures increase the amount of confidence decreases. These results demonstrate the same relationship as *Conf*; lower confidence reduces trading. The only other significant predictor of volume is *Liti*. *Liti* is positively related to volume with a total effect of 0.2541.

Table VI contains results for tests on *Vix* as the dependent variable. Confidence is positively related to the Vix index as evidenced by the results for *Conf* and *Weak*. *Conf* carries a positive net effect and *Weak* carries a negative net effect. This indicates that confidence and *Vix* are positively related. Although, the aggregated effect of the confidence measures is small, our sentiment measures have strong relationships with future levels of *Vix*. *SentLM* is significant and negatively related to future volatility. Negative speech is again the driver behind the relationship. *Neg* is statistically significant and positively related to future volatility.

Table V: Predicting Volume Using All Tweets

This table contains results for Granger-Causality tests where the language measures use all tweets in the sample. All variables have been standardized with mean zero and unit variance. Dependent variable, *Y*, is the daily volume of the S&P 500. Coefficients on *Y* are denoted as α_{t-i}. Independent variables, *X*, are our language measures. Coefficients on *X* are denoted as β_{t-i}. Bull is bullishness. Conf and Sent are the DataSift confidence and sentiment measures. Strg, Weak, Unct, Neg, and Pos are scores for strong, weak, uncertainty, negative, and positive language from the Loughran and McDonald dictionary. SentLM is a sentiment score based off of the difference between positive and negative word usage from the Loughran and McDonald Dictionary.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Coefficient	Bull	Conf	Strg	Weak	Unct	Sent	SentLM	Neg	Pos	Liti
α_{t-1}	0.344***	0.296***	0.307***	0.339***	0.351***	0.363***	0.363***	0.369***	0.317***	0.314***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
α_{t-2}	-0.000907	-0.0638	-0.0548	-0.0201	-0.0250	-0.0462	-0.0375	-0.0338	-0.0309	-0.0407
	(0.991)	(0.429)	(0.514)	(0.807)	(0.763)	(0.583)	(0.652)	(0.689)	(0.710)	(0.622)
α_{t-3}	-0.0101	-0.00415	-0.0128	0.0549	0.0563	0.0165	0.0124	0.0158	0.0243	0.0390
	(0.905)	(0.959)	(0.878)	(0.499)	(0.495)	(0.844)	(0.881)	(0.851)	(0.770)	(0.634)
α_{t-4}	-0.147*	-0.166**	-0.102	-0.159*	-0.180**	-0.121	-0.113	-0.126	-0.157*	-0.0903
	(0.084)	(0.042)	(0.222)	(0.057)	(0.034)	(0.149)	(0.177)	(0.142)	(0.063)	(0.262)
α_{t-5}	0.0578	-0.0173	0.00698	0.0148	0.00957	0.0215	0.0467	0.0392	-0.000257	0.0400

	(0.489)	(0.831)	(0.932)	(0.856)	(0.908)	(0.790)	(0.565)	(0.641)	(0.998)	(0.611)
β_{t-1}	0.0484	0.170**	-0.0454	-0.123	-0.162**	0.166**	0.108	-0.116	-0.0213	-0.0124
	(0.536)	(0.024)	(0.567)	(0.161)	(0.036)	(0.033)	(0.165)	(0.148)	(0.783)	(0.864)
β_{t-2}	0.0287	0.0700	-0.0775	-0.0220	-0.0278	-0.00913	-0.0115	0.0179	0.00806	0.229***
	(0.720)	(0.364)	(0.335)	(0.803)	(0.724)	(0.911)	(0.884)	(0.827)	(0.917)	(0.002)
β_{t-3}	-0.139*	0.149*	-0.155*	-0.167*	-0.111	-0.0248	-0.0311	0.0170	-0.0161	0.0785
	(0.087)	(0.057)	(0.051)	(0.059)	(0.153)	(0.766)	(0.692)	(0.832)	(0.834)	(0.296)
β_{t-4}	-0.0155	-0.0300	0.0798	0.135	0.106	0.0302	0.112	-0.0582	0.136*	-0.0803
	(0.848)	(0.708)	(0.324)	(0.123)	(0.175)	(0.723)	(0.158)	(0.475)	(0.076)	(0.287)
β_{t-5}	0.0666	0.0105	-0.0640	0.130	0.115	-0.0848	-0.0872	0.110	0.0941	0.0393
	(0.397)	(0.895)	(0.427)	(0.135)	(0.138)	(0.317)	(0.266)	(0.167)	(0.221)	(0.597)
Sum of β_{t-i}	-0.0108	0.3695	-0.2621	-0.047	-0.0798	0.07747	0.0902	-0.0293	0.20076	0.2541
Prob > F	0.4291	0.0742	0.1718	0.0899	0.0588	0.3099	0.4725	0.5286	0.3923	0.0202
N	181	181	181	181	181	181	181	181	181	181

Table VI: Predicting Vix Using All Tweets

This table contains results for Granger-Causality tests where the language measures use all tweets in the sample. All variables have been standardized with mean zero and unit variance. Dependent variable, *Y*, is the daily closing value of the Vix Index. Coefficients on *Y* are denoted as α_{t-i}. Independent variables, *X*, are our language measures. Coefficients on *X* are denoted as β_{t-i}. Bull is bullishness. Conf and Sent are the DataSift confidence and sentiment measures. Strg, Weak, Unct, Neg, and Pos are scores for strong, weak, uncertainty, negative, and positive language from the Loughran and McDonald dictionary. SentLM is a sentiment score based off of the difference between positive and negative word usage from the Loughran and McDonald Dictionary.

Coefficient	(1) Bull	(2) Conf	(3) Strg	(4) Weak	(5) Unct	(6) Sent	(7) SentLM	(8) Neg	(9) Pos	(10) Liti
α_{t-1}	0.704***	0.750***	0.672***	0.650***	0.671***	0.685***	0.657***	0.650***	0.690***	0.666***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
α_{t-2}	0.151	0.100	0.157*	0.194**	0.186**	0.157*	0.177*	0.181**	0.174*	0.159*
	(0.105)	(0.294)	(0.092)	(0.034)	(0.044)	(0.093)	(0.053)	(0.046)	(0.062)	(0.085)

α_{t-3}	0.0935	0.0480	0.0777	0.114	0.0983	0.0928	0.0834	0.0755	0.0790	0.0907
	(0.317)	(0.616)	(0.406)	(0.215)	(0.287)	(0.321)	(0.364)	(0.409)	(0.399)	(0.329)
α_{t-4}	-0.0745	0.0462	-0.0567	-0.0905	-0.0715	-0.0668	-0.0642	-0.0540	-0.0694	-0.0745
	(0.421)	(0.629)	(0.540)	(0.319)	(0.437)	(0.471)	(0.479)	(0.548)	(0.453)	(0.418)
α_{t-5}	0.00284	-0.0724	0.0138	-0.00965	0.00505	0.00692	-0.0115	-0.0118	0.00657	0.000845
	(0.971)	(0.356)	(0.859)	(0.900)	(0.948)	(0.929)	(0.880)	(0.875)	(0.932)	(0.991)
β_{t-1}	-0.000310	0.106***	-0.00666	-0.0794*	-0.0275	-0.0404	-0.0304	0.0518	0.0675	0.0139
	(0.994)	(0.008)	(0.874)	(0.090)	(0.755)	(0.337)	(0.455)	(0.194)	(0.427)	(0.728)
β_{t-2}	-0.0205	-0.0164	-0.0628	-0.0878*	-0.109	0.0378	0.0362	-0.0537	-0.117	0.0338
	(0.627)	(0.700)	(0.145)	(0.063)	(0.220)	(0.390)	(0.386)	(0.195)	(0.174)	(0.402)
β_{t-3}	0.0431	-0.0331	-0.00425	-0.00798	-0.0131	-0.0463	-0.0532	0.0578	0.00777	0.0559
	(0.313)	(0.441)	(0.921)	(0.867)	(0.882)	(0.299)	(0.201)	(0.163)	(0.928)	(0.171)
β_{t-4}	-0.0677	0.0954**	-0.0518	0.0680	0.183**	-0.0455	-0.107**	0.120**	0.119	0.0331
	(0.111)	(0.031)	(0.228)	(0.146)	(0.040)	(0.320)	(0.011)	(0.005)	(0.165)	(0.420)
β_{t-5}	0.0295	-0.100**	-0.0308	0.0870*	0.101	-0.0132	-0.0204	0.0351	0.0627	0.0409
	(0.476)	(0.025)	(0.475)	(0.066)	(0.261)	(0.772)	(0.630)	(0.409)	(0.470)	(0.315)
Sum of β_{t-i}	-0.0159	0.0519	-0.1563	-0.02018	0.1344	-0.1076	-0.1748	0.211	0.13967	0.1776
Prob > F	0.6764	0.0063	0.3486	0.0389	0.1578	0.5139	0.0338	0.0037	0.4253	0.4035
N	181	181	181	181	181	181	181	181	181	181

Across Tables IV thru VI there are three common findings. First, *Bull* is never significant. This could be related to two issues. First, the poor performance of the Twitter bullishness measure to predict daily returns could indicate that bullishness is being incorporated more quickly into prices and a more appropriate test would be over shorter intraday time intervals. Second, there are a relatively small amount of tweets that are categorized as buy, sell, or hold. Therefore, bullishness sentiment is poorly captured by Twitter. An alternative explanation is that users do not feel they have sufficient length in a tweet to express an opinion on buying or selling behavior. A large number of tweets state that to get analysis about a particular stock one must follow the hyperlink in the tweet. So, the tweet does not reveal any opinion, but rather guides people to longer form

articles that may reveal opinion. We argue that it may be beneficial to not only analyze the content of each tweet but also analyze the content of linked webpages.

The second common finding is that the significance of our sentiment measures which are based on the Loughran and McDonald dictionary tend to have lower p-values than those based on the DataSift sentiment measure. It would appear that even on an informal communication channel such as Twitter, compensating for specific financial language is important.

The third common finding is that the significant lagged values often occur between days t_{-2} and t_{-5}. This seems to indicate that confidence and sentiment in Twitter may not immediately incorporate itself into prices, volume, and volatility. Or it may imply some persistent effect of confidence and sentiment. We do not believe that this necessarily represents inefficiencies in the market. Tests of intraday period likely reveal quick response to the information in tweets. The longer horizons in our test are likely capturing some other behavior.

4.2. User Gender

We extract one characteristic of Twitter users and examine its effect on the predictive ability of tweets. We have chosen to examine the gender of the user because of strong psychological evidence which suggests there are differences between men and women in this area.

Table VII contains results for Granger-Causality tests using only tweets made by males where *Ret* is the dependent variable. We find that eight of our ten language measures significantly predict returns. *Conf* is more statistically significant and of similar economic magnitude for males when compared to the general population. *Weak* and *Unct* are significant but lose some of their strength relative to the same tests in Table IV. *Sent*, *SentLM*, *Neg*, *Pos*, and *Liti* are statistically significant and economically significant.

Table VIII shows results for tests on *Vol*. Of our confidence language measures only *Unct* has any predictive ability. Its cumulative effect is positive, which is opposite to the results for Table V. This is interesting given that men are expected to be more overconfident than women. So, men's confidence is not highly impactful on aggregate trading; and, trading increases when there is more uncertainty language from men.

Three of our sentiment language measures for men predict volume. *Sent* has an aggregate negative effect and is economically large. Consistent with *Sent*, *Neg* and *Liti* both have positive net effects on volume. So, as sentiment becomes more positive for men, aggregate trading goes down. Aggregate trading could not be explained by the total sentiment measures in Table V.

Table VII: Predicting Return Using Male Tweets

This table contains results for Granger-Causality tests where the language measures use only tweets by males in the sample. All variables have been standardized with mean zero and unit variance. Dependent variable, *Y*, is the daily return of the S&P 500. Coefficients on *Y* are denoted as α_{t-i}. Independent variables, *X*, are our language measures. Coefficients on *X* are denoted as β_{t-i}. Bull is bullishness. Conf and Sent are the DataSift confidence and sentiment measures. Strg, Weak, Unct, Neg, and Pos are scores for strong, weak, uncertainty, negative, and positive language from the Loughran and McDonald dictionary. SentLM is a sentiment score based off of the difference between positive and negative word usage from the Loughran and McDonald Dictionary.

Coefficient	(1) Bull	(2) Conf	(3) Strg	(4) Weak	(5) Unct	(6) Sent	(7) SentLM	(8) Neg	(9) Pos	(10) Liti
α_{t-1}	-0.0823	-0.0366	-0.0952	-0.108	-0.0863	-0.109	-0.103	-0.0981	-0.0562	-0.115
	(0.282)	(0.626)	(0.209)	(0.152)	(0.254)	(0.151)	(0.176)	(0.195)	(0.460)	(0.129)
α_{t-2}	-0.0647	-0.0659	-0.0724	-0.0629	-0.0502	-0.0613	-0.0292	-0.0444	-0.0545	0.0896
	(0.395)	(0.376)	(0.335)	(0.400)	(0.504)	(0.414)	(0.697)	(0.553)	(0.475)	(0.237)
α_{t-3}	-0.0166	-0.0707	-0.0306	-0.00276	-0.00408	-0.0298	-0.0413	-0.0630	-0.0359	0.0422
	(0.829)	(0.345)	(0.686)	(0.970)	(0.956)	(0.690)	(0.574)	(0.386)	(0.634)	(0.576)
α_{t-4}	-0.134*	-0.0573	-0.132*	-0.112	-0.111	-0.145*	-0.124*	-0.128*	-0.113	0.136*
	(0.084)	(0.448)	(0.083)	(0.133)	(0.136)	(0.054)	(0.093)	(0.079)	(0.138)	(0.073)
α_{t-5}	-0.161**	-0.212***	-0.169**	-0.167**	-0.163**	-0.195**	-0.174**	-0.185**	-0.168**	-0.179**
	(0.039)	(0.005)	(0.028)	(0.026)	(0.031)	(0.011)	(0.020)	(0.012)	(0.027)	(0.018)
β_{t-1}	0.0435	-0.250***	0.0153	0.127	0.0149	0.0538	0.0853	-0.146*	-0.179**	0.136**
	(0.593)	(0.002)	(0.851)	(0.104)	(0.845)	(0.480)	(0.280)	(0.062)	(0.048)	(0.077)
β_{t-2}	0.0286	0.245***	0.0497	0.202**	0.190**	-0.102	-0.196**	0.227**	0.198**	-0.0590
	(0.736)	(0.003)	(0.550)	(0.011)	(0.015)	(0.183)	(0.019)	(0.006)	(0.032)	(0.458)
β_{t-3}	-0.0185	-0.0981	0.0219	-0.0542	-0.0986	0.140*	0.180**	-0.209**	-0.114	-0.115
	(0.829)	(0.242)	(0.790)	(0.499)	(0.207)	(0.069)	(0.034)	(0.013)	(0.225)	(0.163)

β_{t-4}	0.118	-0.177**	0.101	-0.0553	-0.0827	0.141*	0.190**	-0.171**	-0.0778	-0.0867
	(0.164)	(0.033)	(0.224)	(0.490)	(0.292)	(0.070)	(0.027)	(0.045)	(0.412)	(0.295)
β_{t-5}	-0.0276	0.0514	0.0817	-0.148*	-0.0988	0.106	0.0183	-0.0494	-0.00896	-0.0458
	(0.736)	(0.529)	(0.322)	(0.062)	(0.206)	(0.175)	(0.825)	(0.555)	(0.927)	(0.573)
Sum of β_{t-i}	0.144	-0.2287	0.2696	0.0715	-0.0752	0.3388	0.2776	-0.3484	-0.18176	-0.44256
Prob > F	0.7545	0.0006	0.4034	0.0176	0.0662	0.0266	0.0023	0.0002	0.0765	0.0563
N	181	181	181	181	181	181	181	181	181	181

Table VIII: Predicting Volume Using Male Tweets

This table contains results for Granger-Causality tests where the language measures use only tweets by males in the sample. All variables have been standardized with mean zero and unit variance. Dependent variable, *Y*, is the daily volume of the S&P 500. Coefficients on *Y* are denoted as α_{t-i}. Independent variables, *X*, are our language measures. Coefficients on *X* are denoted as β_{t-i}. Bull is bullishness. Conf and Sent are the DataSift confidence and sentiment measures. Strg, Weak, Unct, Neg, and Pos are scores for strong, weak, uncertainty, negative, and positive language from the Loughran and McDonald dictionary. SentLM is a sentiment score based off of the difference between positive and negative word usage from the Loughran and McDonald Dictionary.

Coefficient	(1) Bull	(2) Conf	(3) Strg	(4) Weak	(5) Unct	(6) Sent	(7) SentLM	(8) Neg	(9) Pos	(10) Liti
α_{t-1}	0.332***	0.318***	0.305***	0.338***	0.331***	0.363***	0.369***	0.350***	0.273***	0.280***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
α_{t-2}	-0.00505	-0.0360	-0.0518	-0.0288	-0.0277	-0.0580	-0.0381	-0.0541	-0.0746	-0.0732
	(0.953)	(0.664)	(0.538)	(0.729)	(0.742)	(0.495)	(0.651)	(0.526)	(0.355)	(0.366)
α_{t-3}	-0.0129	0.0244	-0.00879	0.0347	0.0292	-0.0161	0.00175	-0.0177	-0.0249	-0.00291
	(0.880)	(0.769)	(0.917)	(0.675)	(0.728)	(0.849)	(0.983)	(0.836)	(0.759)	(0.971)
α_{t-4}	-0.163*	-0.130	-0.131	-0.148*	-0.160*	-0.100	-0.101	-0.128	-0.200**	-0.0836
	(0.057)	(0.119)	(0.121)	(0.079)	(0.060)	(0.236)	(0.232)	(0.136)	(0.014)	(0.284)
α_{t-5}	0.0345	0.0201	0.0281	0.0132	-0.0209	-0.000788	0.0294	0.00117	-0.0190	0.0438

	(0.681)	(0.805)	(0.734)	(0.873)	(0.803)	(0.992)	(0.719)	(0.989)	(0.814)	(0.568)
β_{t-1}	0.0381	0.0448	-0.0391	-0.114	-0.109	0.137*	0.0976	-0.0853	-0.0952	-0.0223
	(0.641)	(0.579)	(0.634)	(0.150)	(0.164)	(0.074)	(0.234)	(0.308)	(0.271)	(0.761)
β_{t-2}	0.0327	-0.00660	-0.0302	0.0261	0.0166	-0.0623	-0.0227	0.0605	0.102	0.253***
	(0.701)	(0.936)	(0.715)	(0.746)	(0.833)	(0.422)	(0.790)	(0.485)	(0.248)	(0.001)
β_{t-3}	-0.152*	0.00505	-0.130	-0.0414	0.0000741	-0.112	-0.0557	0.0849	0.0943	0.162**
	(0.076)	(0.951)	(0.115)	(0.602)	(0.999)	(0.149)	(0.513)	(0.324)	(0.291)	(0.041)
β_{t-4}	-0.0691	-0.0221	0.00842	0.0881	0.119	0.00823	0.118	-0.0227	0.210**	-0.0381
	(0.420)	(0.786)	(0.919)	(0.269)	(0.130)	(0.916)	(0.169)	(0.793)	(0.020)	(0.633)
β_{t-5}	0.0137	0.0820	-0.0467	0.155**	0.208***	-0.172**	-0.149*	0.184**	0.184**	0.111
	(0.868)	(0.305)	(0.569)	(0.049)	(0.008)	(0.027)	(0.070)	(0.028)	(0.046)	(0.158)
Sum of β_{t-i}	-0.1366	0.10315	-0.1618	0.1138	0.23467	-0.20107	-0.0118	0.2214	0.4951	0.4656
Prob > F	0.3767	0.8944	0.466	0.172	0.0299	0.0468	0.3937	0.1523	0.0012	0.0003
N	181	181	181	181	181	181	181	181	181	181

Table IX: Predicting Vix Using Male Tweets

This table contains results for Granger-Causality tests where the language measures use only tweets by males in the sample. All variables have been standardized with mean zero and unit variance. Dependent variable, *Y*, is the daily closing value of the Vix Index. Coefficients on *Y* are denoted as α_{t-i}. Independent variables, *X*, are our language measures. Coefficients on *X* are denoted as β_{t-i}. Bull is bullishness. Conf and Sent are the DataSift confidence and sentiment measures. Strg, Weak, Unct, Neg, and Pos are scores for strong, weak, uncertainty, negative, and positive language from the Loughran and McDonald dictionary. SentLM is a sentiment score based off of the difference between positive and negative word usage from the Loughran and McDonald Dictionary.

Coefficient	(1) Bull	(2) Conf	(3) Strg	(4) Weak	(5) Unct	(6) Sent	(7) SentLM	(8) Neg	(9) Pos	(10) Liti
α_{t-1}	0.701***	0.752***	0.666***	0.658***	0.666***	0.651***	0.636***	0.647***	0.703***	0.667***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
α_{t-2}	0.149	0.101	0.158*	0.184**	0.182**	0.167*	0.176*	0.169*	0.151	0.160*

	(0.109)	(0.293)	(0.090)	(0.044)	(0.049)	(0.068)	(0.051)	(0.062)	(0.111)	(0.083)
α_{t-3}	0.0932	0.0485	0.0790	0.121	0.107	0.0895	0.0796	0.0745	0.0788	0.0818
	(0.316)	(0.612)	(0.400)	(0.189)	(0.248)	(0.330)	(0.380)	(0.415)	(0.409)	(0.378)
α_{t-4}	-0.0803	0.0568	-0.0663	-0.0821	-0.0693	-0.0770	-0.0650	-0.0519	-0.0471	-0.0673
	(0.384)	(0.556)	(0.477)	(0.364)	(0.447)	(0.397)	(0.468)	(0.563)	(0.616)	(0.463)
α_{t-5}	0.00909	-0.0893	0.0148	0.0001890	0.00275	-0.000818	-0.0161	-0.0143	-0.00750	-0.0118
	(0.907)	(0.253)	(0.849)	(0.998)	(0.971)	(0.991)	(0.830)	(0.848)	(0.924)	(0.879)
β_{t-1}	-0.00420	0.121***	-0.0125	-0.0496	-0.0163	-0.0233	-0.0493	0.0553	0.0895	0.0169
	(0.921)	(0.005)	(0.771)	(0.233)	(0.848)	(0.564)	(0.238)	(0.186)	(0.381)	(0.683)
β_{t-2}	-0.0316	-0.0962**	-0.0349	-0.0881**	-0.132	0.00708	0.0408	-0.0542	-0.146	0.0803*
	(0.477)	(0.031)	(0.431)	(0.038)	(0.125)	(0.862)	(0.353)	(0.218)	(0.159)	(0.059)
β_{t-3}	0.0435	0.00731	-0.0151	0.0339	0.0603	-0.0446	-0.0432	0.0498	0.0557	0.0103
	(0.329)	(0.869)	(0.732)	(0.426)	(0.485)	(0.273)	(0.325)	(0.257)	(0.595)	(0.815)
β_{t-4}	-0.0855*	0.112**	-0.0414	0.0529	0.179**	-0.0888**	-0.124***	0.121**	0.134	0.0603
	(0.055)	(0.012)	(0.347)	(0.213)	(0.040)	(0.031)	(0.005)	(0.007)	(0.203)	(0.172)
β_{t-5}	0.0206	-0.0752*	-0.0531	0.0774*	0.112	-0.0421	-0.0377	0.0366	-0.00926	0.0142
	(0.632)	(0.084)	(0.225)	(0.069)	(0.203)	(0.311)	(0.392)	(0.411)	(0.932)	(0.743)
Sum of β_{t-i}	-0.0572	0.0689	-0.157	0.0265	0.203	-0.1917	-0.2134	0.2085	0.12394	0.182
Prob > F	0.5099	0.0043	0.3506	0.0407	0.0819	0.0906	0.0046	0.0027	0.4589	0.199
N	181	181	181	181	181	181	181	181	181	181

Table IX tests the predictive ability of tweets by men on *Vix*. Overall the predictive ability of men's communication is stronger than that of all Twitter users. Tests of confidence related measures show that men's confidence as measured by *Conf* shows a positive relationship between confidence and volatility. However, *Weak* and *Unct* are also positively related to volatility. These signs are opposite of what we might expect the relationship to be. The difference cannot be resolved without knowing the proprietary method used by DataSift for their measure.

Men's sentiment related language is highly predictive of *Vix*. *Sent* and *SentLM* are negative and significant with net effects of about -0.20. Men's negative word usage as measured by *Neg* is also significant and economically large. These results are similar to using all tweets.

Table X contains results for Granger-Causality tests using only tweets made by females. Five of our language measures have significance for predicting *Ret*. The magnitude of the effect of *Conf* is greater for women than it is for men. *Unct* carries a positive net effect whereas for men it was negative. A positive coefficient on *Unct* is consistent with a negative coefficient on *Conf*. We anticipate that women are less confident and are more likely to use weak language, therefore the stronger results for our confidence related measures are interesting.

Women's sentiment is positively related to returns. Negative word usage is again the driver behind the significant relationship for women. *Neg* carries a net effect of -0.1258. Litigious language is marginally significant for women and has a negative sign which is consistent with *Liti* being another type of negative communication.

Table X: Predicting Return Using Female Tweets

This table contains results for Granger-Causality tests where the language measures use only tweets by females in the sample. All variables have been standardized with mean zero and unit variance. Dependent variable, *Y*, is the daily return of the S&P 500. Coefficients on *Y* are denoted as α_{t-i}. Independent variables, *X*, are our language measures. Coefficients on *X* are denoted as β_{t-i}. Bull is bullishness. Conf and Sent are the DataSift confidence and sentiment measures. Strg, Weak, Unct, Neg, and Pos are scores for strong, weak, uncertainty, negative, and positive language from the Loughran and McDonald dictionary. SentLM is a sentiment score based off of the difference between positive and negative word usage from the Loughran and McDonald Dictionary.

Coefficient	(1) Bull	(2) Conf	(3) Strg	(4) Weak	(5) Unct	(6) Sent	(7) SentLM	(8) Neg	(9) Pos	(10) Liti
α_{t-1}	-0.0656	-0.0434	-0.0767	-0.127*	-0.104	-0.0665	-0.0802	-0.0770	-0.0809	-0.0788
	(0.388)	(0.566)	(0.313)	(0.093)	(0.170)	(0.381)	(0.291)	(0.313)	(0.290)	(0.302)
α_{t-2}	-0.0664	-0.0777	-0.0646	-0.0958	-0.0484	-0.0644	-0.0258	-0.00901	-0.0494	-0.0655
	(0.379)	(0.300)	(0.391)	(0.201)	(0.517)	(0.398)	(0.730)	(0.904)	(0.517)	(0.389)
α_{t-3}	0.000403	-0.0528	-0.0234	-0.0288	0.0106	0.00455	-0.00729	-0.0285	-0.00752	-0.0257
	(0.996)	(0.485)	(0.758)	(0.703)	(0.887)	(0.953)	(0.922)	(0.699)	(0.922)	(0.732)
α_{t-4}	-0.120	-0.0553	-0.112	-0.145*	-0.124*	-0.108	-0.100	-0.105	-0.119	-0.134*
	(0.117)	(0.466)	(0.142)	(0.055)	(0.099)	(0.156)	(0.174)	(0.152)	(0.124)	(0.077)
α_{t-5}	-0.142*	-0.150**	-0.157**	-0.177**	-0.172**	-0.151*	-0.161**	-0.151**	-0.156**	-0.166**
	(0.068)	(0.043)	(0.041)	(0.019)	(0.023)	(0.051)	(0.031)	(0.041)	(0.045)	(0.031)
β_{t-1}	-0.0234	-0.230***	0.00683	0.219***	0.106	0.0698	-0.00921	-0.0256	-0.0603	-0.0162
	(0.762)	(0.002)	(0.929)	(0.004)	(0.157)	(0.389)	(0.902)	(0.727)	(0.432)	(0.830)

β_{t-2}	0.0587	0.00625	0.0749	0.149*	0.157**	-0.140*	-0.171**	0.204***	0.0359	0.0395
	(0.455)	(0.934)	(0.334)	(0.055)	(0.039)	(0.092)	(0.024)	(0.006)	(0.643)	(0.600)
β_{t-3}	-0.0456	-0.0664	-0.0565	0.105	0.115	0.0536	0.144*	-0.157**	0.00589	-0.200***
	(0.563)	(0.381)	(0.459)	(0.180)	(0.134)	(0.514)	(0.062)	(0.039)	(0.940)	(0.010)
β_{t-4}	0.0646	-0.140*	0.115	-0.0304	-0.120	-0.0241	0.172**	-0.184**	-0.000392	-0.0456
	(0.413)	(0.068)	(0.135)	(0.694)	(0.116)	(0.777)	(0.027)	(0.017)	(0.996)	(0.557)
β_{t-5}	-0.106	0.150*	0.0662	-0.0974	-0.0803	0.0778	-0.0264	0.00368	-0.0335	-0.0398
	(0.173)	(0.054)	(0.390)	(0.207)	(0.294)	(0.349)	(0.735)	(0.962)	(0.666)	(0.607)
Sum of β_{t-i}	-0.0517	-0.28015	0.20643	0.3452	0.1777	0.0371	0.10939	-0.1258	0.00061	-0.2621
Prob > F	0.7175	0.0033	0.4322	0.0047	0.032	0.5598	0.0262	0.0047	0.9625	0.1322
N	181	181	181	181	181	181	181	181	181	181

Table XI shows results for predicting *Vol* using women's tweets. For women only two language measures predict *Vol*, weak and litigious word usage. The net effect of *Weak* on *Vol* is -0.12298. This is larger than the same test using all tweets or male tweets and the sign is opposite. A negative sign on *Weak* is consistent with the idea that confidence and trading are positively related. Tests using *Liti* are significant with a positive coefficient indicating that aggregate trading increases as women are more negative.

Table XII contains results for *Vix* using women's tweets. Of confidence measures, *Conf* is positive and significant. *Weak* is significant and economically large. *Weak* has a net effect of -0.0846, consistent with *Conf*. *Unct* is also statistically significant but economically small. Sentiment measure *SentLM* is marginally significant and negatively related to future volatility. Negative word measure *Neg* appears to be driving the result as it is the only other sentiment measure with significance. It is highly significant and has a net effect of 0.1712.

Overall the evidence suggests that men's language has more predictive power explaining *Ret* than the language of women; however, using all tweets to predict *Ret* performs best. Men's tweets are stronger predictors of future volume than women's tweets. *Vix* is more related to language by men, although there is predictive ability for both genders. These results complement prior literature that show men trade more than women therefore it is not unusual to expect that men's sentiment is the stronger determinant of market performance over short horizons. Our sample contains more than 70% of tweets by men, also strengthening the results towards them.

The implication that *Ret*, *Vol*, or *Vix* can be predictable with Twitter data may lead to a variety of trading strategies. Although we do not explore trading strategies here, we note that hedge funds use data and analysis of this nature to determine trading activity.

Table XI: Predicting Volume Using Female Tweets

This table contains results for Granger-Causality tests where the language measures use only tweets by females in the sample. All variables have been standardized with mean zero and unit variance. Dependent variable, *Y*, is the daily volume of the S&P 500. Coefficients on *Y* are denoted as α_{t-i}. Independent variables, *X*, are our language measures. Coefficients on *X* are denoted as β_{t-i}. Bull is bullishness. Conf and Sent are the DataSift confidence and sentiment measures. Strg, Weak, Unct, Neg, and Pos are scores for strong, weak, uncertainty, negative, and positive language from the Loughran and McDonald dictionary. SentLM is a sentiment score based off of the difference between positive and negative word usage from the Loughran and McDonald Dictionary.

Coefficient	(1) Bull	(2) Conf	(3) Strg	(4) Weak	(5) Unct	(6) Sent	(7) SentLM	(8) Neg	(9) Pos	(10) Liti
α_{t-1}	0.342***	0.318***	0.316***	0.333***	0.346***	0.344***	0.353***	0.346***	0.317***	0.329***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
α_{t-2}	-0.0185	-0.0531	-0.0452	-0.0279	-0.0477	-0.0418	-0.0475	-0.0466	-0.0444	-0.0411
	(0.825)	(0.517)	(0.589)	(0.732)	(0.564)	(0.616)	(0.566)	(0.577)	(0.589)	(0.617)
α_{t-3}	0.000743	0.0167	-0.000517	0.0515	0.0593	0.0251	0.0310	0.0288	0.0181	0.0330
	(0.993)	(0.838)	(0.995)	(0.521)	(0.472)	(0.763)	(0.708)	(0.730)	(0.826)	(0.685)
α_{t-4}	-0.137	-0.138*	-0.0971	-0.139*	-0.150*	-0.108	-0.122	-0.127	-0.128	-0.0970
	(0.100)	(0.094)	(0.245)	(0.087)	(0.071)	(0.196)	(0.145)	(0.134)	(0.122)	(0.231)
α_{t-5}	0.0553	0.0309	0.0316	0.0393	0.0365	0.0313	0.0400	0.0419	0.0277	0.0276
	(0.502)	(0.699)	(0.698)	(0.619)	(0.650)	(0.699)	(0.620)	(0.614)	(0.734)	(0.726)
β_{t-1}	0.0533	0.127*	-0.0177	-0.0941	-0.129*	0.0350	0.0869	-0.0431	0.0680	0.0000632
	(0.485)	(0.084)	(0.817)	(0.209)	(0.080)	(0.658)	(0.251)	(0.573)	(0.360)	(0.999)
β_{t-2}	-0.0142	0.0402	-0.0872	0.0000231	0.000784	-0.00474	-0.0461	0.0553	0.0295	0.204***
	(0.853)	(0.586)	(0.253)	(1.000)	(0.992)	(0.953)	(0.544)	(0.467)	(0.693)	(0.005)
β_{t-3}	-0.126	0.116	-0.125*	-0.204***	-0.121	0.000249	0.0195	-0.0171	0.0146	0.0771
	(0.104)	(0.119)	(0.099)	(0.007)	(0.109)	(0.998)	(0.798)	(0.820)	(0.846)	(0.301)
β_{t-4}	0.0249	-0.00816	0.0779	0.0341	0.0536	0.0915	0.0878	-0.0481	0.0663	-0.142*
	(0.749)	(0.913)	(0.306)	(0.649)	(0.473)	(0.267)	(0.247)	(0.526)	(0.377)	(0.056)
β_{t-5}	0.0688	-0.0172	-0.0170	0.141*	0.0683	-0.0844	-0.0869	0.0801	0.00796	0.0248

	(0.366)	(0.817)	(0.823)	(0.062)	(0.360)	(0.296)	(0.252)	(0.290)	(0.915)	(0.739)
Sum of β_{t-i}	0.00671	0.25784	-0.169	-0.12298	-0.12732	0.03761	0.0612	0.0271	0.18636	0.16396
Prob > F	0.5445	0.3734	0.3764	0.0349	0.1895	0.7987	0.5928	0.8106	0.8254	0.0353
N	181	181	181	181	181	181	181	181	181	181

Table XII: Predicting Vix Using Female Tweets

This table contains results for Granger-Causality tests where the language measures use only tweets by females in the sample. All variables have been standardized with mean zero and unit variance. Dependent variable, *Y*, is the daily closing value of the Vix Index. Coefficients on *Y* are denoted as α_{t-i}. Independent variables, *X*, are our language measures. Coefficients on *X* are denoted as β_{t-i}. Bull is bullishness. Conf and Sent are the DataSift confidence and sentiment measures. Strg, Weak, Unct, Neg, and Pos are scores for strong, weak, uncertainty, negative, and positive language from the Loughran and McDonald dictionary. SentLM is a sentiment score based off of the difference between positive and negative word usage from the Loughran and McDonald Dictionary.

Coefficient	(1) Bull	(2) Conf	(3) Strg	(4) Weak	(5) Unct	(6) Sent	(7) SentLM	(8) Neg	(9) Pos	(10) Liti
α_{t-1}	0.719***	0.739***	0.690***	0.638***	0.659***	0.710***	0.663***	0.665***	0.691***	0.682***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
α_{t-2}	0.130	0.0959	0.156*	0.195**	0.197**	0.132	0.169*	0.182**	0.175*	0.163*
	(0.163)	(0.314)	(0.097)	(0.032)	(0.031)	(0.164)	(0.066)	(0.048)	(0.063)	(0.081)
α_{t-3}	0.110	0.0579	0.0726	0.118	0.104	0.106	0.0960	0.0788	0.0743	0.0782
	(0.238)	(0.546)	(0.444)	(0.193)	(0.248)	(0.262)	(0.299)	(0.394)	(0.433)	(0.404)
α_{t-4}	-0.0941	0.0515	-0.0451	-0.0899	-0.0851	-0.0821	-0.0642	-0.0569	-0.0719	-0.0778
	(0.309)	(0.591)	(0.628)	(0.316)	(0.344)	(0.381)	(0.485)	(0.530)	(0.443)	(0.400)
α_{t-5}	0.0121	-0.0729	0.000950	-0.00805	0.0000723	0.0131	-0.0149	-0.0177	0.00672	0.0152
	(0.877)	(0.346)	(0.990)	(0.915)	(0.999)	(0.867)	(0.846)	(0.814)	(0.932)	(0.846)
β_{t-1}	0.0324	0.125***	-0.0122	-0.0982**	-0.0898	-0.0955**	-0.0229	0.0379	0.0283	-0.00749
	(0.420)	(0.002)	(0.765)	(0.016)	(0.286)	(0.025)	(0.571)	(0.332)	(0.738)	(0.852)
β_{t-2}	-0.0379	0.0127	-0.0681*	-0.0857**	-0.124	0.0657	0.0315	-0.0673*	-0.128	0.00533
	(0.358)	(0.751)	(0.098)	(0.037)	(0.141)	(0.134)	(0.437)	(0.088)	(0.134)	(0.894)

β_{t-3}	0.0438	-0.0166	0.0256	-0.0424	-0.133	0.00399	-0.0468	0.0521	-0.0155	0.0787*
	(0.291)	(0.682)	(0.531)	(0.311)	(0.119)	(0.927)	(0.250)	(0.186)	(0.858)	(0.056)
β_{t-4}	-0.0676	0.0704*	-0.0266	0.0617	0.195**	0.0206	-0.0891**	0.105***	0.0363	0.0208
	(0.104)	(0.085)	(0.513)	(0.136)	(0.022)	(0.646)	(0.030)	(0.009)	(0.672)	(0.611)
β_{t-5}	0.0515	-0.102**	-0.0334	0.0800*	0.161*	-0.0451	-0.0324	0.0435	0.0123	0.0226
	(0.208)	(0.013)	(0.410)	(0.056)	(0.064)	(0.301)	(0.432)	(0.286)	(0.886)	(0.583)
Sum of β_{t-i}	0.0222	0.0895	-0.1147	-0.0846	0.0092	-0.0144	-0.1597	0.1712	-0.0666	0.16791
Prob > F	0.386	0.0022	0.4555	0.0048	0.0142	0.2297	0.107	0.0123	0.7852	0.4386
N	181	181	181	181	181	181	181	181	181	181

4.3. Differences in Confidence and Sentiment

Due to the differences that exist in the way communication by men and women explain returns, volume, and volatility, we look to see if there are differences between the values of our language measures for men and women. There is psychological, economic, and other evidence that suggests there are differences in their communication.

We analyze the differences between genders using t-tests and Wilcoxon-Mann-Whitney rank-sum z-scores. Differences are calculated as the level for women minus the level for men. In Panel A of Table XIII we show differences in language usage between men and women. The confidence measures reveal men are more confident than women in total confidence and use more strong language as measured by *Conf* and *Strg*, respectively. However, men also use more weak and uncertain language than women; this result seems to contradict the prior result. The strong statistical significance of these results is muted by the small practical differences in the values. So, while the differences are not substantively large, they do appear to drive the differences in predictive ability displayed in Tables VII and XII.

Panel B of Table XIII compares our sentiment measures. Men are less bullish than women. Women display higher sentiment than men when measured by *Sent*. Men's sentiment is higher when measured by *SentLM*. Women use more negative and positive language than men. Differences are highly statistically significant but remain practically small. Women also use more litigious language. Women dominate sentiment related language, using stronger language in positive and negative directions. Overall, women are more optimistic than men when they communicate about stocks given these results.

Although the differences in each language measure across both panels are highly statistically significant, the practical differences remain small. However, despite the differences' small scale we conclude there is a meaningful difference in the language of men and women when they communicate about stocks. This conclusion is accentuated by the differences in each gender's ability to predict the market.

Table XIII: Tests of Differences in Sentiment

This table contains results for t-tests and Wilcoxon-Mann-Whitney rank-sum z-scores for differences in our language measure. Differences are calculated as the level for women minus the level for men. Bull is bullishness. Conf and Sent are the DataSift confidence and sentiment measures. Strg, Weak, Unct, Neg, and Pos are scores for strong, weak, uncertainty, negative, and positive language from the Loughran and McDonald dictionary. SentLM is the sentiment score based off of the Loughran and McDonald Dictionaries.

Var	Difference	T-Stat	P-Value	Z-Score	P-Value
Panel A - Confidence					
Conf	-1.4024	-7.8698	0.0000	-76.659	0.0000
Strg	-0.0751	-15.0390	0.0000	-44.669	0.0000
Weak	-0.0333	-5.2439	0.0000	-15.718	0.0000
Unct	-0.0788	-6.5466	0.0000	-25.243	0.0000
Panel B - Sentiment					
Bull	0.1270	16.0200	0.0000	8.3308	0.0000
Sent	0.0209	0.5647	0.5730	0.7676	0.4427
SentLM	-0.0832	-1.8622	0.0632	-0.3327	0.7394
Neg	0.2450	6.3812	0.0000	3.1148	0.0018
Pos	0.1614	5.0725	0.0000	4.9846	0.0000
Liti	0.1140	8.9540	0.0000	5.9913	0.0000

5. Conclusion

In this paper, we use a unique set of Twitter data to examine the relationship between social media communication and the stock market. Twitter is a useful source to gauge investor sentiment and opinion because it is a direct measure of individuals' communication. Our results suggest that daily return, volume, and expected volatility of the S&P 500 can be predicted using different language measures. Specifically, we show market returns may be predicted using confidence and sentiment levels. Volume is mainly predicted by confidence. Expected volatility is most related to sentiment. We examine gender as one important dimension of Twitter user characteristics. Our results show that men are more confident and less optimistic than women when they communicate about stocks. We find differences in the ability of communications by men and women to predict market returns, volume, and volatility. Our findings are consistent with prior psychology, finance, and economics literature.

This paper adds to the debate surrounding the relationship between individual communication and the stock market. This debate had centered around whether individual communication is noise, opinion, or information. We show that individual communication, as captured by Twitter, is unlikely to be noise. The opinion contained within tweets does have some dynamic causal relations with the market. Further study is required to identify if tweets are just opinion or if they contain information in the more traditional sense.

We find that Twitter may not be a good source of gauging bullish or bearish sentiment. This could be related to two main issues. First, the poor performance of the Twitter bullishness measure to predict daily returns could indicate that bullishness is being incorporated more quickly and a

more appropriate test would be over shorter intraday time intervals. Second, this may be due to the small amount of tweets that directly express bullish or bearish sentiment that is distinct from positive or negative sentiment. Future research may overcome these issues by using data from Stocktwits, a social network that specializes in microblogging just about stocks.

Our results have several broader implications for future finance research. The ability to predict returns, volume, and volatility could be constructed into trading strategies. These trading strategies are already in use by institutional and algorithmic traders. This raises the question of why institutions are paying attention to individual investor's sentiment when individuals make up an ever decreasing portion of market volume. There are questions about how Twitter and social media information should be tested within the efficient market hypothesis.

REFERENCES

Agnew, Julie R, 2006, Do behavioral biases vary across individuals? Evidence from individual level 401 (k) data, *Journal of Financial and Quantitative Analysis* 41, 939–962.

Agnew, Julie R, Pierluigi Balduzzi, and Annika E Sunden, 2003, Portfolio choice and trading in a large 401 (k) plan, *American Economic Review* 93, 193–215.

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, 1259–1294.

Asem, Ebenezer, and Gloria Y Tian, 2010, Market dynamics and momentum profits, *Journal of Financial and Quantitative Analysis* 45, 1549–1562.

Baker, Malcolm, and Jeffery Wurgler, 2007, Investor Sentiment in the Stock Market, *The Journal of Economic Perspectives* 21, 129–151.

Barber, Brad M, and Terrance Odean, 2001, Boys will be Boys: Gender, Overconfidence, and Common Stock Investment, *Journal of Banking and Finance* 116, 261–292.

Barsky, Robert B, F Thomas Juster, Miles S Kimball, and Matthew D Shapiro, 1997, Preference Parameters and Behavioral Heterogeneity: An Experimental Approach in the Health and Retirement Study, *The Quarterly Journal of Economics* pp. 537–579.

Bird, Steven, Ewan Klein, and Edward Loper, 2009, *Natural Language Processing with Python - Steven Bird, Ewan Klein, Edward Loper - Google Books* (O'Reilly Media).

Bollen, J, Huina Mao, and Xiaojun Zeng, 2011, Twitter mood predicts the stock market, *Journal of Computational Science* 2, 1–8.

Chisholm, E, and T G Kolda, 1999, New term weighting formulas for the vector space method in information retrieval, Discussion Paper, Oak Ridge National Laboratory ORNL-TM-13756 Oak Ridge, TN.

Chou, Robin K, and Yun-Yi Wang, 2011, A test of the different implications of the overconfidence and disposition hypotheses, *Journal of Banking and Finance* 35, 2037–2046.

Costa, Paul Jr, and Robert R McCrae, 1992, Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual (Psychological Assessment Resources: Odessa, FL).

Costa, Paul Jr, Antonio Terracciano, and Robert R McCrae, 2001, Gender differences in personality traits across cultures: Robust and surprising findings., *Journal of Personality and Social Psychology* 81, 322–331.

Da, Z, J Engelberg, and Pengjie Gao, 2011, In search of attention, *The Journal of Finance* 66, 1461–1499.

Das, Sanjiv R, and Mike Y Chen, 2007, Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web, *Management Science* 53, 1375–1388.

Friedman, Nir, Dan Geiger, and Moises Goldszmidt, 1997, Bayesian network classifiers, *Machine learning* 29, 131–163.

Gonzalez-Bailon, Sandra, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno, 2012, Assessing the Bias in Communication Networks Sampled from Twitter (December 4, 2012),

Halko, Marja-Liisa, Markku Kaustia, and Elias Alanko, 2012, The gender effect in risky asset holdings, *Journal of Economic Behavior & Organization* 83, 66–81.

Hanke, Michael, and Florian Hauser, 2008, On the effects of stock spam e-mails, *Journal of Financial Markets* 11, 57–83.

Hofstede, G H, 1998, *Masculinity and Femininity: The Taboo Dimension of National Cultures* (SAGE Publications, Inc.).

Lewellen, Wilbur G, Ronald C Lease, and Gary G Schlarbaum, 1977, Patterns of investment strategy and behavior among individual investors, *The Journal of Business* 50, 296–333.

Lin, Y C, and Priya Raghubir, 2005, Gender Differences in Unrealistic Optimism About Marriage and Divorce: Are Men More Optimistic and Women More Realistic? , *Personality and Social Psychology Bulletin* 31, 198–207.

Loughran, Tim, and Bill McDonald, 2011, When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, *The Journal of Finance* 66, 35–65.

Lundeberg, M A, and P W Fox, 1994, Highly confident but wrong: Gender differences and similarities in confidence judgments, *Journal of Educational Psychology*.

Lusardi, Annamaria, and Olivia S Mitchell, 2008, Planning and Financial Literacy: How Do Women Fare? , *The American Economic Review* 98, 413–417.

Mizrach, Bruce, and Susan Weerts, 2009, Experts online: An analysis of trading activity in a public Internet chat room, *Journal of Economic Behavior & Organization* 70, 266–281.

Nosek, Brian A., Mahzarin R. Banaji, and Anthony G. Greenwald, 2002, Math = male, me = female, therefore math ≠ me., *Journal of Personality and Social Psychology* 83, 44–59.

Sabherwal, Sanjiv, Salil K Sarkar, and Ying Zhang, 2011, Do Internet Stock Message Boards Influence Trading? Evidence from Heavily Discussed Stocks with No Fundamental News, *Journal of Business Finance & Accounting* 38, 1209–1237.

Säve-Söderbergh, Jenny, 2012, Self-Directed Pensions: Gender, Risk, and Portfolio Choices, *The Scandinavian Journal of Economics* 114, 705–728.

Schmader, Toni, 2002, Gender Identification Moderates Stereotype Threat Effects on Women's Math Performance, *Journal of Experimental Social Psychology* 38, 194–201.

Schubert, Renate, Martin Brown, Matthias Gysler, and Hans Wolfgang Brachinger, 1999, Financial decision-making: are women really more risk-averse? , *The American Economic Review* 89, 381–385.

Statman, Meir, Steven Thorley, and Keith Vorkink, 2006, Investor Overconfidence and Trading Volume, *Review of Financial Studies* 19, 1531–1565.

Steele, Jennifer, Jacquelyn James, and Rosalind Barnett, 2002, Learning in a Man's World: Examining the Perceptions of Undergraduate Women in Male-Dominated Academic Areas, *Psychology of Women Quarterly* 26, 46–50.

Sunden, Annika E, and Brian J Surette, 1998, Gender differences in the allocation of assets in retirement savings plans, *The American Economic Review* 88, 207–211.

Thelwall, Mike, David Wilkinson, and Sukhvinder Uppal, 2009, Data mining emotion in social network communication: Gender differences in MySpace, *Journal of the American Society for Information Science and Technology* 61, 190–199.

Varnhagen, Connie K, G Peggy McFall, Nicole Pugh, Lisa Routledge, Heather Sumida-MacDonald, and Trudy E Kwong, 2009, lol: new language and spelling in instant messaging, *Reading and Writing* 23, 719–733.

Wu, Xindong, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, Phillip S Yu, Zhi-Hua Zhou, Michael Steinbach, David J Hand, and Dan Steinberg, 2008, Top 10 algorithms in data mining, *Knowledge and Information Systems* 14, 1–37.

Zhang, Ying, and Peggy E Swanson, 2008, Are day traders bias free? —evidence from internet stock message boards, *Journal of Economics and Finance* 34, 96–112.