Electron Markets (2017) 27:283–296
DOI 10.1007/s12525-017-0254-5

RESEARCH PAPER

Market sentiment dispersion and its effects on stock return and volatility

Eric. W. K. See-To[1] · Yang Yang[1]

Received: 4 July 2016 / Accepted: 15 March 2017 / Published online: 20 April 2017

Abstract Behavioral economics has revealed that investor sentiment can profoundly affect individual behavior and decision-making. Recently, the question is no longer whether investor sentiment affects stock market valuation, but how to directly measure investor sentiment and quantify its effects. Before the era of big data, research uses proxies as a mediator to indirectly measure investor sentiment, which has proved elusive due to insufficient data points. In addition, most of extant sentiment analysis studies focus on institutional investors instead of individual investors. This is despite the fact that United States individual investors have been holding around 50% of the stock market in direct stock investments. In order to overcome difficulties in measuring sentiment and endorse the importance of individual investors, we examine the role of individual sentiment dispersion in stock market. In particular, we investigate whether sentiment dispersion contains information about future stock returns and realized volatility. Leveraging on development of big data and recent advances in data and text mining techniques, we capture 1,170,414 data points from Twitter and used a text mining method to extract sentiment and applied both linear regression and Support Vector Regression; found that individual sentiment dispersion contains information about stock realized volatility, and can be used to increase the prediction accuracy. We expect our results contribute to extant theories of electronic market financial behavior by directly measuring the individual sentiment dispersion; raising a new perspective to assess the impact of investor opinion on stock market; and recommending a supplementary investing approach using user-generated content.

Keywords Investor sentiment · Text mining · Return and volatility predictability

JEL Classification C55 · C53 · C52

Introduction

The idea of investor sentiment dates back to mid-twentieth when Keynes (1936) proposed that markets are influenced by investors' "animal spirits", causing prices to deviate from fundamentals. This idea is formalized by De Long et al. (1990), who theoretically demonstrated that sentiment changes can lead to noise trading and excessive volatility. Now, the question is no longer whether investor sentiment affects stock market valuation, but how to directly measure investor sentiment and quantify its effects. Extant studies identified two kinds of sentiment measures (Lee et al. 1991; Neal and Wheatley 1998; Brown and Cliff 2004). The first sentiment measures are derived from surveys while the second measures relied on objective variables that correlate with investor sentiment. Both of these two measurements has limitations as they heavily relied on some proxies as mediators to reflect investor sentiment (Schmeling 2009; Carlin et al. 2014). The way to directly measure investor sentiment is still waiting for further exploration.

Despite the number of published works on the issue of investor sentiment, several avenues of research remain unexplored. In particular, the empirical question of a relationship between individual sentiment dispersion and stock price valuation remains unresolved. Dispersion in investor sentiment are often mentioned as a factor that could explain the stock

Responsible Editor: Eric Ngai

✉ Eric. W. K. See-To
eric.see-to@polyu.edu.hk

[1] The Hong Kong Polytechnic University, Hong Kong, Hong Kong

volatility but rarely analyzed (De Long and Shleifer 1991; Shiller 2000). Although various proxies of sentiment are utilized in these papers, most of these research ignored the difference of opinions among investors, or the diversity of investor sentiment, which in theory has impact on asset price, risk and returns (Varian 1985; Qian 2014; Miller 1977) and can lead to higher uncertainty and more trading (Carlin et al. 2014). The fact that sentiment diversity is ignored could result from the proxies of measuring investor sentiment. Some of the methods such as survey questionnaire is difficult to generate enough data points to estimate the sentiment dispersion, some of them can only provide one sentiment score for each week or each month, and common sentiment analysis tools only predict the polarity of the sentiment (i.e. only predict positive or negative for each sentence). The high difficulty of sentiment measurement causes big challenges for scholars. This condition leads to a second limitation. Most of extant research only focuses on institutional investor sentiment (Gao and Kling 2008; Verma and Soydemir 2009), and the individual sentiment is commonly measured by proxies such as the survey of expectation (Brown and Cliff 2004; Fisher and Statman 2000) and consumer confidence (Schmeling 2009). This is despite the fact that United States individual investors have been holding around 50% of the stock market in direct stock investments. Therefore, the role of individual sentiment dispersion in theories of financial behavior and the relationship between this factor and stock price valuation are two pieces of white papers.

To conclude, there are two major research gaps in extant studies of investor sentiment. First, most of extant research only focuses on institutional investor sentiment (Gao and Kling 2008; Verma and Soydemir 2009) and ignores individual sentiment, while individual investors have been holding around 50% of the stock market in direct stock investments. Second, extant studies face big challenges in measuring individual investor sentiment. Most of the previous studies use proxies or questionnaires to conduct a measurement. Such self-report and indirect measurement can be biased. It is also not practical to generate sufficient data points to measure sentiment dispersion on a reasonable time window, such as daily or weekly.

Corresponding to previous research gaps, the objective of this study is clear. We would like to directly measure individual investors' sentiment and to explore the roles of individual sentiment and its dispersion in the stock market. Fortunately, proliferation of online social media and the phenomenal growth of data mining technologies have brought to us the era of big data. Today's digital environment provides previously unavailable measures of investor sentiment. In particular, emerging social platforms such as Twitter and Facebook can potentially provide real-time information on individual sentiment. Clearly, the availability of measures of investor sentiment is only going to increase as we move further into the digital age. The big data era provides us with a great opportunity to overcome difficulties in measuring individual sentiment, and quantifying its effect on stock price valuation.

Consonant with this trend, this study explores the roles of individual sentiment and its dispersion in the stock market. In particular, we provide empirical evidence to show the impact of individual investors' sentiment diversity on stock returns and volatility. Three months' messages of twitter containing stock ticker symbols (e.g. $AAPL, $GOOG) are used to calculate the investor sentiment on each day. The dollar sign tag is for twitter users to express opinions on publicly traded companies. This new tag was popularized by StockTwits, which is an online social media platform for investors and traders to exchange ideas. Labeling each tweet with sentiment scores gives us a direct measure of individual investor sentiment. The sentiment dispersion on each day will be used to predict future stock returns and future realized volatility.

To overcome the difficulties in measuring individual sentiment, we utilize Naive Bayes probability model. Specifically, we use Naive Bayes to assign a probability (p) to each tweet, so that p represents the probability that a tweet is generated from positive sentiment, and 1-p would be the probability of this tweet being negative, given the words it uses. The sentiment score will then be a continuous number ranging from 0 to 1, allowing us to directly measure the dispersion on each day by calculating standard deviation.

To our knowledge, this is one of the earliest study measuring the impacts of individual investors' sentiment dispersion on future stock returns and volatility. We expect this study can fill in the research gap by quantifying sentiment dispersion of individual investor and explore how this factor affects stock return and volatility. More broadly, our work offers the following two contributions. First, the direct measurement of individual sentiment dispersion is an essential building block to further advance theories of financial behavior. Results of this study provide important empirical evidence to extant theories and may inform scholars on the future roadmap for subsequent research. In practice, any improvement in stock volatility and return prediction accuracy helps investors and traders make better informed decisions.

Literature review

The role of investor sentiment in stock market

There has been an increasing interest in using investor sentiment to predict market behaviors, especially stock returns. For example, Baker and Wurgler (2006) studied the investor sentiment and the cross-section stock return, and they discovered that stock returns following a low sentiment period are relatively higher for stocks that are difficult to arbitrage. Tetlock (2007) first used the pessimism information extracted from media content as a proxy to investor sentiment, and he found that the pessimism in media content has the ability to explain the Dow Jones returns. Schmeling (2009) also investigated the relationship between consumer confidences and expected stock returns in 18 industrialized countries, and concluded that sentiment can be used to forecast aggregate market returns. A systematic review done by Akter and Wamba (2016) had addressed the importance of sentiment analysis in different types of markets.

Investor sentiment is also proven to have predictive power on stock movement. Bollen et al. (2011) used twitter to measure public mood to forecast the daily movement of Dow Jones Industrial Average (DJIA), and reported a noteworthy accuracy. Zhang et al. (2011) investigated business engagement through extracting sentiment from tweets and brought insight to the analytics of social networks and online word-of-mouth message diffusion patterns. Similarly, Bing et al. (2014) used twitter sentiment to forecast the stock movement for 30 companies in New York Stock Exchange (NYSE) and NASDAQ, and achieved 76.12% accuracy. In these cases, the prediction is normally about up and down trend.

Currently, there are mixed results about whether investor sentiment is correlated with stock returns and volatility, although in theory trading based on investor sentiment will cause excessive volatility (De Long et al. 1990). Tetlock (2007) only find a weak correlation between his pessimism measure and market volatility. Wang et al. (2006) deems that previous works may have overestimated the predictability of investor sentiment to stock volatility. When he added past volatility into the equation as independent variables, he found that investor sentiment does not predict the future volatility. His research also shows that the past returns and volatility causes the sentiment changes. We try to test this issue as well. We used Vector Auto Regression (VAR) and considered past 5-day data as control variables to avoid overestimation.

Influence of individual investor sentiment

Apart from institutional investors, individual investors, or retail investors are also an important part of the market. The behavior of institutional investors and individual investors in stock market differ in several ways. Stoffman (2008) states that the biggest difference between institutional and individual investors is the fact that institutions exhibit the ability to produce risk-adjusted excess returns. This ability required large amounts of resources. They have sufficient information obtained from consultant, advisor or investment team who is running the portfolio. Compared to institutional investors, individual investors perform much more randomly and irrationally. They do not have the professional knowledge to conduct systematic market analysis and are highly influenced by external environment. As Stoffman (2008) shows, they tend to buy or sell stocks when those stocks are in the public discussion. This attention-based buying and selling can lead individual investors to trade too speculatively and has the potential to influence the pricing and volatility of stocks. Poteshman (2001) documents that individual investors exhibit the same pattern of under-reaction or over-reaction to public information that has been found in stock markets. The irrational decision of individual investor may lead to poor performance of trading activities. Further confirmation comes from Barber et al. (2009b), they show that stocks bought by institutions (sold by individuals) earn strong returns, while stocks bought by individuals (sold by institutions) perform poorly. Above statements are supported by a growing literature. Individual investors often base their investment decisions on their own research, or on the suggestions from media (De Long et al. 1990). Black (1986) believes that these investors sometimes treat noise as information and act on it. Verma and Verma (2007) investigated the sentiments of both institutional and individual investors, and concluded that individual investor sentiments are more irrational than sentiments of institutional investors.

The behavior of individual investors is worth studying because individual investors are proven to be influential in stock market. The model of De Long et al. (1990) states that when noise traders act as a group, they could influence stock price, driving it away from equilibrium. Barber et al. (2009a) then found that individual investors do act as a group and affect stock price. Fisher and Statman (2000) also found evidence that there is a negative relationship between individual investor sentiment and future stock returns.

Impact of opinion difference

Although the role of sentiment level in stock market has been studied by lots of scholars, and it has been proven to have predictive power on stock movement, its derivative, sentiment dispersion is rarely analyzed. Some of the information about volatility may be contained in sentiment dispersion instead of sentiment level. De long et al. (1990) argues that excessive volatility attributes to the unpredictability of noise traders' beliefs. Some of the information about volatility may be contained in sentiment dispersion instead of sentiment level. De long et al. (1990) argues that excessive volatility attributes to the unpredictability of noise traders' beliefs. Unpredictability may increase when people hold various expectations on future returns. The risk may, in turn, drives up the divergence of the dispersion of opinions (Miller 1977). This statement is theoretically supported by Gruca et al. (2005), in that study, they demonstrated that dispersion of the traders' individual forecasts performs well in market prediction. Some empirical evidence has been found showing how opinion differences of institutional investors affect the stock returns, such as Diether et al. (2002), who found that higher dispersion in analysts' forecasts was correlated with lower future returns, and Carlin et al. (2014), who instead found expected return to be negatively correlated to sentiment dispersion. Carlin et al. (2014) also found that sentiment dispersion of institutional investors was negatively associated with return volatility, which is consistent with De Long et al. (1990).

Sentiment measure for individual investors

Even though extant literature has theoretically proved that individual sentiment is significantly related to stock price, the measurement methodologies of this variable are still waiting for further exploration. It is not easy to measure individual investor's sentiment. Past research has used various proxies to tackle different research problems. Schmeling (2009) used consumer confidence as a proxy because his study involves several countries. Brown and Cliff (2004), Fisher and Statman (2000), and Verma and Verma (2008) used the survey data from American Association of Individual Investor (AAII). AAII conduct weekly survey on AAII members each week, asking them about the expectation of market movement direction (Bullish, Neutral, or Bearish). Participants are randomly chosen from 100,000 AAII members, and since this survey targets at individual investors, this result can be a robust measure of individual investors' sentiment. There are two major limitations of using AAII survey. First, result of survey only updates weeks, in which excess volatility caused by noise traders in short run may disappear (Da et al. 2015). Second, for a time series study, this survey can only provide one data point every week, the amount of data points is not sufficient to conduct unbiased evaluation or real-time forecasting.

Besides, previous studies pay more attention to institutional investor's sentiment instead of individual sentiment (Gao and Kling 2008; Verma and Soydemir 2009). Carlin et al. (2014) measured the differences between forecasts among mortgage dealers on Wall Street, and used it as a proxy for sentiment dispersion. But this difference only reflects the sentiment dispersion for institutional investors. This condition stems from the high difficulty in collecting individual data and finding a proper method to conduct analysis.

Fortunately, the development of big data techniques and social media provide possible solutions for above challenges. Das and Chen (2007) developed new algorithm to extract sentiment from talks on stock message boards because of the rapidly increasing volume of data. Bollen et al. (2011) used Opinion Finder and Google-Profile of Mood States on twitter to assign labels to tweets, and achieved a high accuracy in predicting DJIA movement. Rao and Srivastava (2012) used the Naive Bayes classifier to assign positive or negative sentiment on tweets, and they also reported high correlation when correlating with DJIA as well.

Research methodology

In this section, we outline first the data collected for this study. Then the data preparation process is described. Finally, the sentiment analysis methods and correlation analysis methods respectively are discussed. Through the literature review, we have identified two challenges of investigating individual sentiment. The first is the high difficulty in data collection, the second is the tool to conduct an accurate measurement. Leveraging on the development of big data and recent advances in data and text mining techniques, these two challenges can be well solved. Bollen et al. (2011), and Rao and Srivastava (2012) found that twitter contains significant amount of information about stock market, we will also use twitter data to calculate the investor sentiment. Specifically, we only download stock related data (Rao and Srivastava 2012), which provides a better estimation for individual investors.

Stock related tweets

We collect Data for 30 companies in Dow Jones Industrial Average (DJIA). DJIA is a market index comprised of 30 companies which, according to McGraw Hill Financial (n.d.), have an excellent reputation, attract a large number of investors, demonstrate sustainable growth, and are representative of the sector covered by DJIA. During the study period (from Jan 20, 2015 to Jul 17, 2015), one of the DJIA 30 companies, American Multinational Telecommunications Corporation (AT&T), is replaced by Apple Inc. (AAPL) on March 19, 2015. Compared with AT&T, AAPL can provide us with more stock-related tweets. We decide to choose AAPL as one of the DJIA 30 companies instead of AT&T.

Twitter data is collected from Twitter REST API. Messages are filtered with the dollar sign tag ("$"), which is also called the "cashtag" corresponding to the "hashtag" commonly used in twitter to discuss focused topics. The cashtag sign is for investors to share investment related opinions on specific stocks. For example, the message containing "$DIS" will be specifically related to the Disney's stock. This cashtag is popularized by StockTwits, a platform specialized in sharing stock related opinions. There are many recent studies using the "cashtag" filtered tweets to measure investor sentiment. They are selected commonly because they are less noisy than generalized tweets because of its dedication to stock market (Oliveira et al. 2014, 2013a, b).

For the target stocks in DJIA, we collected 1,170,414 tweets from Jan 20, 2015 to Jul 17, 2015. Omitting the weekends and holidays, there are 120 training days in the study period. The average tweet volume ranges from an average of 67 (TRV) to 2540 tweets per day (AAPL), as shown in Fig. 1. We will call this corpus DJIA corpus thereafter, and use this corpus to examine the effect of sentiment dispersion on the stock returns and volatility.

For the same time period, we also collect messages that contain both the cashtag (followed by any stock ticker in New York Stock Exchange or NASDAQ), and the word "bullish" or "bearish". This corpus will serve as a training

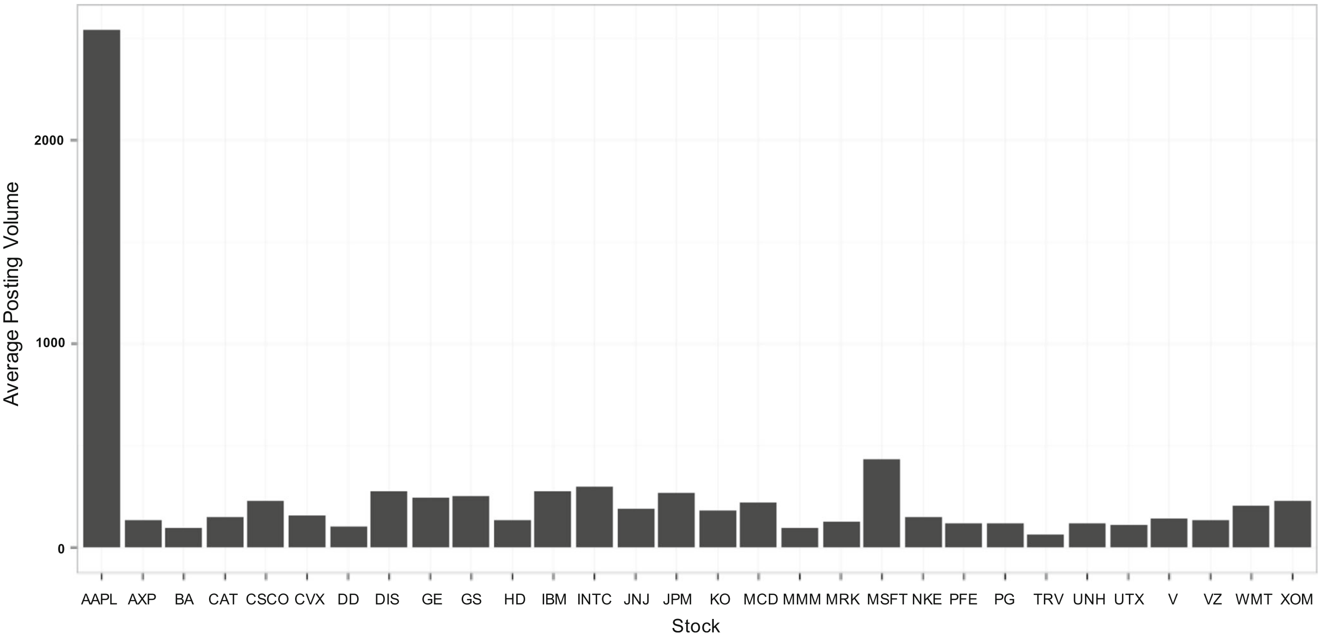

Fig. 1 Average posting volume for DJIA composite stocks

corpus. We simply define the tweets that contain the word "bullish" as bullish in sentiment, meaning the person who posted this twitter expects the price of the target stock to increase, and the tweets that contain the word "bearish" as bearish in sentiment, meaning the person expects the stock price to decrease in the future. This approach of labeling tweets as bullish and bearish is originally used in Mao et al. (2011), who found that the sentiment level in the previous 2 days is significantly related to stock returns. This same approach is later used in Oliveira et al. (2013a) to test the relationship between investor sentiment and volatility. In the data selection stage, we found that some stock observers keep posting tweets with only numbers change. Such message could cause the classifier biased towards the words used in these posts. For example, a tweet is "$AAPL crossed my 116.14, 118 next", and then a repetitive tweet, "$AAPL crossed my 116.24, 118 next", was posted the next minute. For this kind of observers, we will filter them out in our datasets. Also, some tweets posted are too short and only contain modal particle or emoticons. For example, a tweet is "$AAPL WOW!". We added this kind of tweets in stop word list and removed them from training corpus as well. After filtering out repetitive tweets and meaningless tweets, the training corpus contains 6709 bearish tweets and 17,807 bullish tweets. Data collected in this study showed that the bullish tweets (17807) are much more than the bearish tweets (6709). The result implied that the stock market is booming. This is a good opportunity for this study to investigate the role of individual sentiment dispersion under the movement of stock market.

For a classification problem, normally we expect that classes evenly represented in the training dataset, a dataset with this property is called a balanced dataset (Kotsiantis et al. 2006). A balanced dataset is more likely to produce unbiased result (Wei and Dunbrack 2013). In our case, the size of bullish class is much larger than bearish class. The training corpus contains 6709 bearish tweets and 17,807 bullish tweets. We then randomly sampled 6709 bullish tweets in order to balance the bullish and bearish classes, and make our training dataset to fulfill the balanced dataset criteria to reduce the chance of producing a biased result if we use all the bullish tweet in the training process.

Stock returns

Stock return measures the profit of investment. We calculate stock returns using the daily adjusted close price. Adjusted close price is the official close price adjusted by splitting and dividend. Stock close prices are collected from *DataStream*, which provides current and historical data on more than 140,000 securities and instruments in the worldwide market. Daily stock returns, stored in percentile, are calculated using the following formula Oliveira et al. (2013b):

$$\text{return}_t = \frac{p_t - p_{t-1}}{p_t} * 100 \qquad (1)$$

Realized volatility

Volatility is a measure of equity risks. A higher volatility indicates that the amplitude of fluctuation of the corresponding stock is higher. There are many ways to measure the volatility of a stock. A simplest measure is the standard deviation of the daily close prices in a sampled period, say a month. Realized volatility is another measure of equity risk. In this study, we select the realized volatility because it is asymptotically unbiased and it adapts quickly (Corsi 2005) since the realized volatility of day t is only dependent on the returns observed in day t, while independent of the information contained in past returns. We can also see this property in the formula below.

The realized volatility, presented in percentile and in yearly value, on day t is calculated using the following formula (Almgren 2009; Areal and Taylor 2002):

$$RV_t = \sqrt{252*\sum_{i=1}^{n} r_{t,i}*100} \tag{2}$$

Where $r_{t,,i}$ is the intraday returns. The square root of 252 normalizes the daily realized volatility to yearly realized volatility. The intraday return is calculated using the following formula:

$$r_{t,i} = \ln\frac{p_t}{p_{t-1}} \tag{3}$$

Existing empirical work on realized covariance usually compute the realized volatilities based on the 5 to 30 min return interval. Taylor and Xu (1997) relied on 5-min returns in the measurement of daily exchange rate volatilities, Schwert (1998) choose 15-min returns to estimate daily stock market volatilities. Corsi (2005) stated that such frequencies are heuristically chosen to avoid the bias and market microstructure effects. Following previous researchers, we choose five minutes as sampling interval (Corsi 2005), and we use the log return here because it is the convention of calculating realized volatility. The 5-min stock prices are obtained from Bloomberg Professional Service, which contains real time and historical data about equities, securities, derivatives, commodities, and foreign exchanges. It also contains analysts' forecasts, news, and other economic data.

Data pre-processing

Punctuations, numbers and hyperlinks are removed first because hyperlinks and punctuations themselves do not add any new information, while numbers are normally different tweet by tweet. Words are then transformed into lower case. Tweets in both DJIA corpus and training corpus are lemmatized to reduce the variety of words. Lemma is the base form of words, for example, the lemma of words "was", "is", "are" is "be", and the lemma of both "goes", "gone" is "go". We use lemmatizer over stemmer because the lemmatizer returns a real word instead of part of the original word, so that the result is more interpretable. We use the WordNet lemmatizer (Pedersen and Banerjee 2011) because of its popularity.

Words that appear in less than 20 documents in the DJIA corpus is also removed. We select 20 as the threshold based on Griffiths and Steyvers (2004), who filtered words which appear in less than 5 documents when the corpus contains 28,154 documents. The abstracts in the corpus used by Griffiths and Steyvers (2004) normally contain around 100 words, the average number of words in the DJIA corpus is 9.2 words. Considering the number of documents and the number of words in each document, we select 20 as threshold. A total number of 204,599 words are reduced to 12,441 in this step. The filtered words are least frequent words and typos.

This paragraph describes a method to filter out domain specific stop words. Stopwords refer to most common words in a language. Stopwords normally contain little information. In this case, stopwords have less distinguishing power between bullish and bearish sentiment. Since the DJIA corpus is related to stock market, and the word used in twitter is slightly different from everyday English, we choose to build corpus specific stopwords list. We create stopwords list based on the Term Frequency - Inverse Document Frequency (TF-IDF) score. There are many variations in calculating TF-IDF score, here we follow the one used in Hornik and Grün (2011). The formula is presented below.

$$\text{tfidf}_t = \frac{1}{|\{d \in D : t \in d\}|} \sum_{d=1}^{D} f_{t,d} * \log \frac{N}{|\{d \in D : t \in d\}|} \tag{4}$$

Where d stands for a single document, D is whole corpus, N is the number of documents in the corpus, and t stands for term. $|\{d \in D : t \in d\}|$ is the number of documents that contain term t. We consider 1% of all tue words (124 words) as stop words. Common stopwords such as "be", "the", "to", "have", "but" etc. all receive low TF-IDF scores. The word "rt", which stands for retweet, is also successfully identified as a stopword. Specially, words such as "stock", "market", "trade", "invest", "company", "nasdaq", "finance", "news", "daytrading", "inc" are all regarded as stopwords, which is desirable because they are common words in the field of finance, and provide little information about sentiment.

We also look at the face validity of building training corpus in this way. Face validity implies that a method that is put into practical use, should appear to be practical (Mosier 1947). We extracted those words which affect the sentiment most to see whether they are interpretable. In the Naïve Bayes model we will introduce in the next section, the influence of words is determined by the odds ratio, that is, the ratio of the occurrence in bullish records and the occurrence in bearish records. Also, if a word appears frequently, the influence will also be larger. So we calculate the influence simply by the product of odds ratio and the word frequency. We then rank the words, and select top 10 words for both bullish and bearish classes. The results are shown as follows (Table 1).

Having looked at the antonyms, we move on to other words in each class. For bullish class, investors use "BO" for "breakout" in short. Breakout is commonly used to describe the point where stock price breaks the level of resistance and keep growing up. When investors use the word, it means the investors observe a breakout, and expect the stock price to continue increasing. "Nice" is just for investors to describe an optimistic situation such as "nice double bottom", "nice monthly pattern", "nice breakout" etc. "Strong" is also a positive word that is used in "strong stocks", "strong bullish trend". It also appears in phrase "strongly bearish", but the frequency is much lower. Investors also use the phrase "bullish flow" much more than "bearish flow". Therefore the word "flow" has a high influence. Finally, "stockaction" has a very high influence because few messages with negative sentiment contain this word. When we search this term in past tweets, the results are generally about positive news and stock recommendations.

For bearish class, it is not apparent why some words are related to bearish market such as "chart", "support", "sentiment", "trend", and "obv". When we look at the training set, we find that people use "chart" more often when describing a bearish trend, and refer to "support level" and "bearish sentiment" when in the description. They also use "OBV" (On Balance Volume) to support their argument. We regard this phenomenon as a convention of word selection. Although there are some words we find hard to interpret, but a large proportion of influential words are reasonable. The method described above successfully identifies words that will be used when experiencing bullish and bearish sentiment. Therefore, we would say using word "bullish" and "bearish" to build training set is reasonable, and the Naïve Bayes classifier based on this training set can be used to assign reasonable sentiment scores for DJIA corpus.

Sentiment level and dispersion

We use Naive Bayes as a classifier. Although this is a simple classifier, it is proven to be effective in Natural Language Processing (NLP). The fundamental theorem behind Naive Bayes classifier is the Bayes Theorem. Specifically, we first assume the prior probability of being bullish and bearish are both 50%. For each word, the classifier will update the probability of bullish and the probability of bearish. Normally, the last step is to choose the class that achieves the highest probability. However, since sentiment has intensity, simply classify a tweet as bullish or bearish may cause information loss. Therefore, here we normalize the probability so that the probability of bullish and bearish sum to one, and treat the bullish probability as the sentiment score. So that a higher sentiment score represents a bullish sentiment, and a lower sentiment represents a bearish sentiment.

Table 1 Top 10 most influential words in both bullish and bearish classes

Bullish	"bullishness", "bullish", "stockaction", "call", "nice", "flow", "bo", "higher", "strong", "long"
Bearish	"bearish", "bearishness", "short", "trend", "lower", "support", "chart", "put", "sentiment", "obv"

Although Naïve Bayes classifier has been proved to be valid in tweets sentiment analysis (Rao and Srivastava 2012), we still want to test if it can produce unbiased result for sentiment measurement in our context. We would like to do a further step to validate this method. As we mentioned above, now we have a tweet corpus with "bullish" and "bearish" label, which contains 6709 bullish tweets and 6709 bearish tweets (thereafter referred as labeled dataset D0). We followed the steps below to conduct the accuracy check. First, we divided D0 into 15 sub-datasets (D1-D15) and randomly chose 10 of them as training sets, 5 of them as test sets. We used the training set to train a Naive Bayes classifier and applied this classifier on the test set to obtain a sentiment score for each tweet. This sentiment score can be used to interpret if the tweet is bullish or bearish. Second, we compared the result given by Naive Bayes classifier and the label assigned by Twitter. If their results aligned with each other, then we say the measurement model is validated. Finally, we redid previous steps for several times to evaluate the overall performance of Naïve Bayes classifier. The overall result is good, at about 75%.

The test result is shown in Table 2.

After the validation process, the result produced by Naïve Bayes classifier was sufficiently reliable. Then we can calculate the sentiment score for each tweet in the DJIA corpus. We group these tweets according to the stocks and dates. For each stock on each day, we can generate a score for aggregate sentiment level (*Sent*) using the mean of sentiment scores, and a score for sentiment dispersion (*Disp*) using the standard deviation of all sentiment scores. In order to ensure the accuracy of these two measures, we only choose the days that contain more than 50 data points (Nash 2001).

Table 2 Accuracy test of Naïve Bayes classifier

Process	Training set	Test set	Accuracy
1	1,3,4,5,7,8,9,12,13,15	2,6,10,11,14	74.23%
2	2,3,4,5,7,8,9,10,12,14	1,6,11,13,15	76.96%
3	3,4,6,7,8,11,12,13,14,15	1,2,5,9,10	75.46%
4	1,2,3,4,5,7,9,10,11,14	6,8,12,13,15	72.89%
5	1,2,3,4,6,7,8,11,13,15	5,9,10,12,14	76.38%
Average			75.18%

Results: Information contained in sentiment dispersion

Since this study investigates the relationship between individual sentiment dispersion and realized volatility, and this topic is not studied by previous scholars. Thus, we cannot compare the results to extant theories to seek support or difference. However, the results can be an essential building block to remedy current research gap in this field and provide a foundation for further study.

Information about realized volatility

We first use multiple regression to test whether there is a linear relationship between sentiment dispersion and stock return and volatility. If there is such relationship, we would also like to know whether the relationship is positive or negative. Multiple regression is useful in this situation. The independent variables are the past 5-day sentiment dispersion (*L5(Disp$_t$)*). In order to eliminate the effects of past returns, volatilities and sentiment, we also include past 5-day returns (*L5(Return$_t$)*), volatility (*L5(Rv$_t$)*), and sentiment level (*L5(Sent$_t$)*) as control variables. To evaluate the effects of sentiment dispersion. We compare the adjusted R^2 of following two equations:

$$RV_t = L5(RV_t) + L5(\mathrm{return}_t) + L5(\mathrm{Sent}_t) \tag{5}$$

$$RV_t = L5(RV_t) + L5(\mathrm{return}_t) + L5(\mathrm{Sent}_t) + L5(\mathrm{Disp}_t) \tag{6}$$

Similarly, we use the same set of independent variables to explain the normalized daily stock returns using the following equations.

$$\mathrm{return}_t = L5(RV_t) + L5(\mathrm{return}_t) + L5(\mathrm{Sent}_t) \tag{7}$$

$$\begin{aligned}\mathrm{return}_t = {} & L5(RV_t) + L5(\mathrm{return}_t) + L5(\mathrm{Sent}_t) \\ & + L5(\mathrm{Disp}_t)\end{aligned} \tag{8}$$

In the multiple regression, adding any new predictor would cause the R^2 to increase, as long as the coefficient of the new predictor is not zero, which rarely happen in reality. Adjusted R^2 adjusts the R^2 by the number of predictors so that a model is punished for having more variables. Therefore, if adjusted R^2 increases, we could say that the sentiment dispersion contains information about the dependent variables. We also applied ANOVA F test to see whether any of the new independent variables are necessary to explain the dependent variable. ANOVA.

We estimate the exploratory and predictive power of sentiment dispersion using pooled regression. As stated in Hsiao (2014), pooling panel data can possibly generate more accurate result in predictions since it provides the possibility of learning from other groups. Pooled Ordinary Least Square (OLS) was used in Tetlock et al. (2008) to forecast companies accounting earnings, and was used in Antweiler and Frank (2004) to predict volatilities of 45 companies. There are some situations where pooled regression is not suitable. Hsiao (2014) stated that pooled OLS could lead to biased result if the intercepts are heterogeneous. We solve this problem by standardizing the return and volatility series, as used in Tetlock et al. (2008). The second situation is when the slopes across groups are different. In this section, we would like to examine the effect of sentiment dispersion on the market level, therefore, we make the assumption that the effect of dispersion is independent on an individual stock.

The pooled OLS result of eq. 6 is presented in the Table 3. We first examine the coefficients of sentiment dispersions, and the significance of these coefficients. Generally speaking, the past sentiment dispersions negatively affect the realized volatility. Past 1 day sentiment dispersion is most significantly related to the current realized volatility, with the confidence larger than 99%. Past 3-day and 4-day sentiment dispersion also negatively affects the realized volatility, at 95% confidence and 90% confidence respectively. Finally, the effect of sentiment dispersion reverses to positive with 90% confidence level.

According Theil and Nagar (1961), The use of an adjusted R^2 is an attempt to eliminate the condition of R^2 automatically increasing when more explanatory variables are added to the model. When the increase in R^2 is not totally adjusted away by using adjusted R^2, this adds an additional dimension of reliability to the results. The results of this study showed that adjusted R^2 increases when sentiment dispersion L5(Dispt) is added in the model, indicating its significant explanatory power for realized volatility. The adjusted R^2 increases from

Table 3 Explaining realized volatility using sentiment dispersion with different time lags

	Equation (5)	Equation (6)	Add current Dispersion
Disp$_1$		−2.027***	-2.459 ***
Disp$_2$		−0.925	−1.155 *
Disp$_3$		−1.460**	−1.638**
Disp$_4$		−1.322*	−1.320*
Disp$_5$		1.120*	0.889
Disp$_0$			1.686 **
Adj. R^2	0.235	0.245	0.247
F-test *p* value		<0.001***	0.019**

*** 99% confidence level; ** 95% confidence level; *90% confidence level

0.245 to 0.247. The increase of 0.01 is comparable with many papers that try to use information to explain stock price (Bollen et al. 2011; Tetlock 2007; Chordia et al. 2002).

This result does not seem to be consistent with the theories of institutional investors' sentiment, which state that the institutional sentiment dispersion positively affects the realized volatility. We suspect, based on Efficient Market Theory (EMT), that the information about sentiment dispersion may have been incorporated into stock price within a day. Therefore, we add the current sentiment dispersion into the equation, and the result is shown in the third column of Table 3. The adjusted R^2 of the new equation increases from 0.245 to 0.247, indicating that the individual sentiment dispersion would affect the realized volatility on the same day. We can see that the current individual sentiment dispersion is positively related to the same day's realized volatility. Intuitively, this result suggests that the individual sentiment dispersion will firstly increase the stocks' realized volatility, and it reverses the effect and starts to decrease the stocks' realized volatility on the following days. The *p* value from ANOVA F-test suggests the same result that past sentiment dispersion adds significant amount of information to the stock price volatility, and the current sentiment dispersion is also reflected in the same day's stock volatility.

Information about returns

The pooled OLS result of Eq. 8 is listed in Table 4. Only the past 2-day sentiment dispersion is significantly related to the daily stock returns, the coefficient is 2.04, and it is significant at 99% confidence level. The adjusted R^2 increases from 0.0271 to 0.0291. The increment is not significant, and the R^2 themselves are quite small. Therefore, we conclude that the past sentiment dispersion does not contain much information about stock returns, and daily stock returns are much harder to predict than realized volatility.

Table 4 Explaining daily stock returns using sentiment dispersion with different time lags

	Equation (7)	Equation (8)	Add current Dispersion
$Disp_1$		−0.186	−0.199
$Disp_2$		2.042***	1.958**
$Disp_3$		0.016	−0.029
$Disp_4$		−1.013	−0.908
$Disp_5$		−1.071	−1.118
$Disp_0$			−0.096
Adj. R^2	0.0271	0.0291	0.0288
F-test *p* value		0.070*	0.899

*** 99% confidence level; ** 95% confidence level; *90% confidence level

We also test whether stock market is so efficient that the information contained in sentiment dispersion is incorporated into stock price within the same day. Therefore, we add the same day's sentiment dispersion as an independent variable, and the result is shown in the last column. First of all, the adjusted R^2 does not increase. Instead, R^2 decreases from 0.0291 to 0.0288, suggesting that the current sentiment dispersion does not add any new information into same days' stock returns. We can draw the same conclusion by the *p* value from ANOVA F-test (0.899).

Predictive power

To validate the result, we separate the study period in to training and testing periods. The training period is used to train the models and the test set is used to assess the performance of these models. We treat the first 75% (dates before Jun 05, 2015) as training data, and the last 25% (dates on or after Jun 05, 2015) as test data. The training data is used to train the classifiers and the test data is used for evaluation. Note that since the analysis in this sections does not depend on any conclusions made in the last section, therefore, this data splitting method will not cause overfitting. The performance is evaluated using the test set. Performance is evaluated using the Root Mean Square Error (RMSE). The formula is:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - Y_i\right)^2}$$

Again, we use the independent variables listed in the Eqs. 5, 6 ,7, 8, and compare the performance of Eqs. 5 and 6, and the performance of eqs. 7 and 8. If sentiment dispersion has the predictive power to stock realized volatility, we would expect the RMSE of Eq. 6 to be less than the RMSE of Eq. 5, and if sentiment dispersion has the predictive power to daily stock returns, we would expect the RMSE of Eq. 8 to be less than that of Eq. 7.

Here we also consider whether the effects of sentiment dispersion are non-linear. The method we use is the Support Vector Regression (SVR). SVR is a regression model of Support Vector Machine (SVM) that was developed by Drucker et al. (1997). Similar to SVM, SVR seeks to minimize its error, and maximize a goal function by fitting a higher dimension hyperplane. It has been extended to solve nonlinear problems, and is proven to be effective in financial time series (Tay and Cao 2002; Kim 2003). We use the same set of inputs as the multiple regression presented in the equation above, and the adjusted R^2 is also used to evaluate whether sentiment dispersion contains any information about the stock return and volatility.

Additionally, we compare the RMSE of all models with three benchmarks models. The first benchmark model is using past 1-day data as the prediction. For example, we will use the yesterday's stock return as the prediction for today. The second benchmark model is using the average of past 5 days' data as predictions. The third benchmark model is using the average of all data in the training data as predictions. If traditional financial theories are correct, that the stock movement is unpredictable, then we would expect the RMSE of any previous model to fall around the RMSE of the benchmark.

The RMSE of predicting stock returns and volatility is shown in Table 5. The first comparison is between benchmarks and all models. For predicting realized volatility, we can see that all models are better than the benchmark models, which means the stock volatility can possibly be predicted using past information. Since we have normalized the realized volatility company by company, there should be no company specific information contained in the past realized volatility. We then examine whether sentiment dispersion could add additional information for predicting stock volatility. When predicting realized volatility, sentiment dispersion causes the RMSE to decrease 0.008 in linear model, and 0.004 in SVR. Although the decrease is not very significant, but both linear model and SVR agree that the sentiment dispersion adds additional information when predicting volatility.

When predicting returns, we can see that average of training set provides a better prediction to the returns in the test set (RMSE =0.816). Which is surprising because many previous studies found that stock returns can be predicted by investor sentiment (Baker and Wurgler 2006; Tetlock 2007; Schmeling 2009). However, if we compare the model performance to the other benchmarks, we will see that the models actually perform better. We also tried to use the Past 2-day to 5-day return as a prediction, and the RMSE is all around 1.15. This means the return series is quite noisy, and using these return series as predictions may confuse the training model. In fact, if we remove the past returns as inputs, we will obtain a better result. The linear model with sentiment dispersion achieves a RMSE of 0.819, linear model without sentiment dispersion achieves a RMSE of 0.820, SVR with sentiment dispersion obtained a RMSE of 0.839, and the RMSE of SVR without sentiment dispersion is 0.847. Again, the performance is not as good as the benchmark 1. We again conclude that the return series is hard to predict using the past realized volatility, sentiment, or sentiment dispersion.

Table 5 RMSE of four models and three benchmarks

Model	Returns	RV
Multiple regression – without dispersion	0.820	0.715
Multiple regression – with dispersion	0.819	0.707
SVR – without dispersion	0.847	0.733
SVR – with dispersion	0.839	0.729
Benchmark 1 - Past 1-day	1.163	0.816
Benchmark 2 - Avg. of past 5 days	0.913	0.792
Benchmark 3 - Avg. of training set	0.816	0.873

We can also see that from the Figures below. Figure 2 shows the relationship between real realized volatility and predictions given by multiple regression with sentiment dispersion. We add a regression line and the 95% confidence interval in the plot. We can see that the prediction is correlated with the real value. This result shows that the realized volatility is predictable. On the other hand, the predicted returns do not seem to have a relationship with the real returns, the significance level is less than 90%, suggesting return series is much harder to predict than realized volatility (Fig. 3).

Robustness tests

We use another measure of sentiment dispersion to confirm the impact of sentiment dispersion. Here we select the Median Absolute Deviation (MAD) to measure the sentiment dispersion. MAD is calculated in three steps. Firstly, median is subtracted from all the values. Secondly, the absolute values are calculated for these differences, and finally, the median is found to represent the dispersion. The advantage of MAD is that it is more robust to extreme values than standard deviation because it uses median instead of mean.

We also perform above experiments using the MAD as the dispersion measure. When predicting the realized volatility, sentiment dispersion shows the same negative effect as before. Past 1-day and 3-day sentiment dispersion show the most significant negative impact on realized volatility. The adjusted R^2 also increase from 0.235 to 0.241. When predicting returns, past 2-day and 5-day sentiment seem to influence the current stock returns. The adjusted R^2 also increases from 0.271 to 0.287. When fitting SVR on the dataset to predict realized volatility, the adjusted R^2 increase from 0.455 to 0.478, and increase from 0.317 to 0.344 when predicting returns.

Finally, when separating the training and the test dataset, we also found a decrease in RMSE when predicting both returns and realized volatility. Again, the decrease is not significant when predicting returns. When sentiment dispersion is included, the RMSE decrease from 0.828 to 0.827 when predicting returns using linear model. SVR sees a decrease of RMSE from 0.877 to 0.861, which is a better decrease, but the RMSE of SVR is worse than the RMSE of linear model. When predicting realized volatility, sentiment dispersion reduced the RMSE from 0.715 to 0.709 when using linear model, and from 0.733 to 0.730 when using SVR.

Discussion and conclusion

In this study, we used both linear regression and SVR to show that sentiment dispersion contains information about the stock

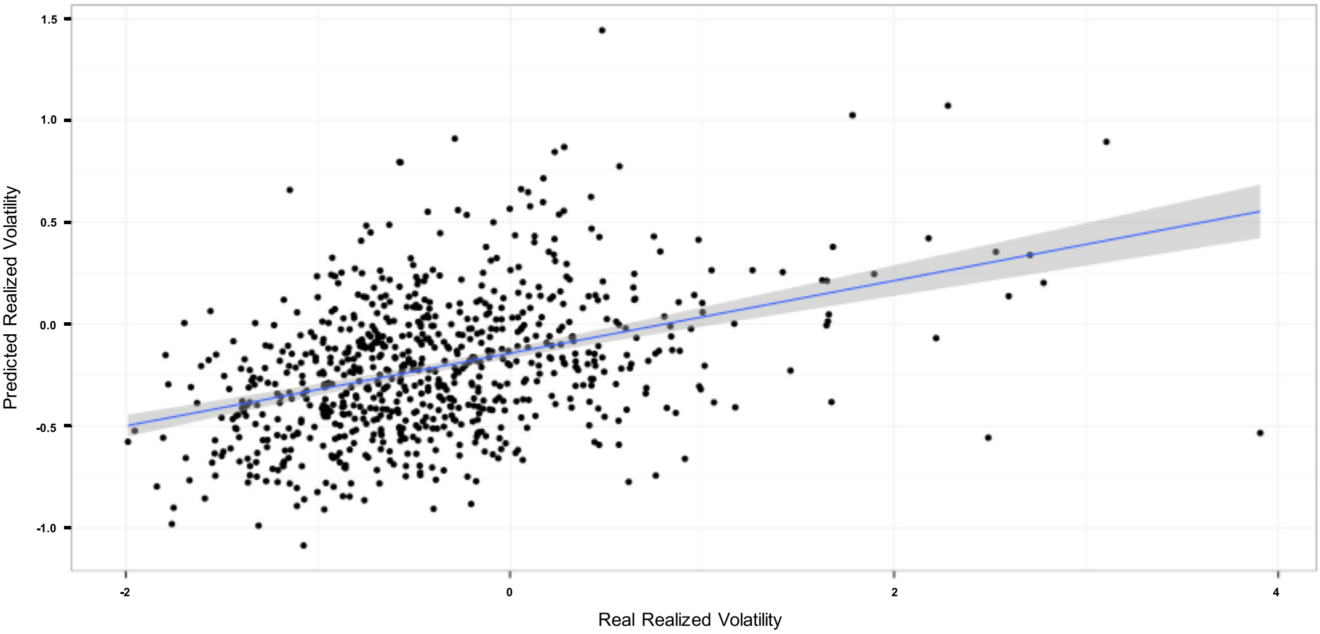

Fig. 2 Relationship between predicted realized volatility and real realized volatility

volatility and stock returns. Specially, sentiment dispersion raises the realized volatility on the same day, and then reduce the realized volatility on the following several days. Subsequent analysis shows that sentiment dispersion can provide additional predictive power to realized volatility. Different from what theory suggests, sentiment dispersion does not contain much information about the daily stock returns. As a robustness test, we also used Median Absolute Deviation to measure sentiment dispersion, and the result confirms our findings. The direct measuring of individual sentiment remedies current research gap and advance theories of financial behavior. The findings uncover the potential predictive power of sentiment dispersion and raise a new perspective to assess the impact of investor opinion on stock market.

We showed the value of our proposed approach in extracting semantic sentiment of stock related tweets and evaluate their predictive power on stock volatility and returns. As literature (Varian 1985; Qian 2014; Miller 1977; Carlin et al. 2014) suggests, sentiment dispersion could contain information about future returns and volatility, and can be used to increase the prediction accuracy. It is practically important to forecast realized volatility because some derivatives (e.g. options) are priced

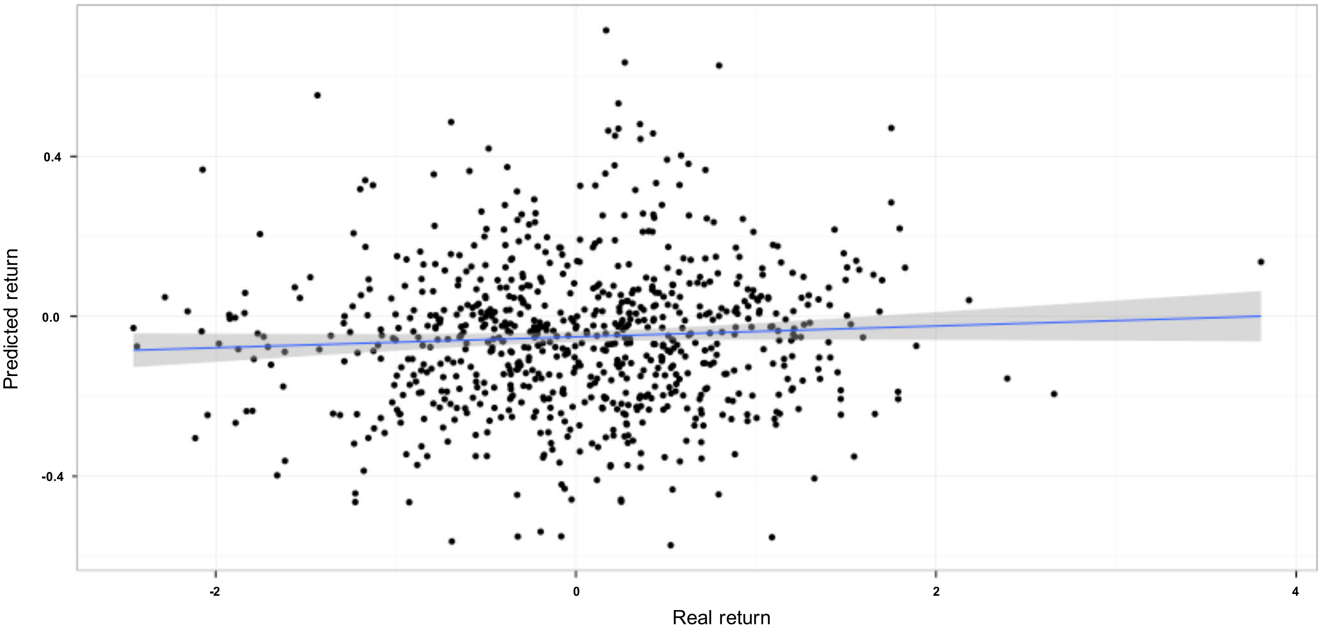

Fig. 3 Relationship between predicted returns and real returns

based on it, and realized volatility acts as input in various finance models. From the results, we can see that all the experiments agree that sentiment dispersion is related to the realized volatility. Particularly, sentiment dispersion will increase the volatility, but the effects happen on the same day, and reduce the realized volatility on the next day. The effects on the stock returns take place more than three days. In addition, the number of tweets is found to be influential to the future stock volatilities and stock returns. Finally, the sentiment dispersion is also affected by the past stock returns and volatility. The effects are showed to be bidirectional, and future research can confirm and address this phenomenon.

This research is one of the earliest attempts to measure the sentiment and its dispersion for individual investors. It was hard to directly measure the sentiment specifically for individual investors in the past. Previous researchers attempted to use some proxies as a mediator to reflect individual investor sentiment. For example, Schmeling (2009) used the consumer confidence level as a proxy for individual investor sentiment. Normally they will obtain only one sentiment score for each day, each week or each month, so that it is hard to confidently measure the sentiment dispersion. Carlin et al. (2014) measured the differences between forecasts among mortgage dealers on Wall Street, and used it as a proxy for sentiment dispersion. But this difference only reflects the sentiment dispersion for institutional investors. In this study, leveraging on the development of twitter and the prevalent use of dollar tag to exchange opinions about stocks, we are able to directly measure the individual investors' sentiment, and thus further measuring the sentiment without any proxies. This successful attempt can significantly advance theories of financial behavior.

Past research is not able to find a significant relationship between individual investor sentiment and stock volatility (Wang et al. 2006; Tetlock 2007). We suggest that information about volatility may be well contained in sentiment dispersion, and this study also assesses sentiment dispersion's ability to forecast future stock price, which can be further used to calculate security risks and predict derivative price. In addition, we examine whether sentiment dispersion of individual investors contains information about stock returns. De Long et al. (1990) suggests that investor sentiment dispersion may increase the risk and uncertainty, thus investors expect higher risk premium and a higher return, while Miller (1977) suggests that difference in investor opinion may in fact decrease the stock return. We provide empirical evidence showing the relationship between individual investors' sentiment dispersion and stock returns. The significant result in this study can increase the accuracy of predicting stock movement.

Last but not the least, the result of this study highlight the role played by sentiment dispersion of investors on the stock market by providing a large amount of empirical evidences. Investors come to the market bringing with them deep-rooted differences that can be traced to their wealth, income, social status and education. These differences affect the way investors approach the market, evaluate the stocks and design their trading strategies. The effect of their sentiment and behavior does not average out in aggregate but directly impacts the market, generating more potential predictors. With the insight of sentiment dispersion, this study could be a roadmap for subsequent scholars.

This study also carries significant managerial implications for the use of social media as a strategic tool to conduct stock price evaluation and prediction. Previous stock market analysis mainly based on company valuation and professional financial measures such as net income analysis or free cash flow calculation. These methods are commonly used by institutional investors. Given the importance of individual sentiment dispersion for stock volatility, social media can become a useful data source for stock valuation. Social media (like Twitter in this study) is a major platform where individual investors express their opinions. Evidence has shown that a large amount of data stream is generated from social media every day (Ekbia et al. 2015), and by the end of 2012, around 2.5 exabytes of data per day were brought about, and the number was doubling every 40 months or so (McAfee and Brynjolfsson 2012). The significant results of this study showed that information generated on social media is valuable in stock market evaluation, especially for volatility prediction. Stock analysts now may consider involving the information generated on social media to achieve more accurate predictive results, with support from the rapid development of social media analytics and big data techniques.

There are two areas for further investigation. In this study, we used linear regression and SVR to examine if the sentiment dispersion contains information about stock volatility and stock returns. These two methods correspond to linear and non-linear relationship respectively. This is an issue that requires further investigation. If time permits, it would be better to use multiple methods to verify the linear and non-linear relationship respectively. Further research should consider using other linear methods, such as hypothesis test, to test if the results align with linear regression. Similarly, using other non-linear methods to test the performance of SVR. If possible, we can consider using the Pearson correlation coefficient to indicate the strength of a linear relationship between sentiment dispersion and stock price volatility.

Second, in this study, we collected data from 30 companies in Dow Jones Industrial Average, which only represents one part of US stock market. The result obtained may differ in other types of stock. Further investigation should be conducted with other stock indices, such as S&P 500, to yield further insights.

References

Akter, S., & Wamba, S. F. (2016). Big data analytics in E-commerce: A systematic review and agenda for future research. *Electronic Markets, 26*(2), 173–194.

Almgren, R. (2009). High frequency volatility. New York University.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance, 59*(3), 1259–1294.

Areal, N. M., & Taylor, S. J. (2002). The realized volatility of FTSE-100 futures prices. *Journal of Futures Markets, 22*(7), 627–648.

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance, 61*(4), 1645–1680.

Barber, B. M., Odean, T., & Zhu, N. (2009a). Do retail trades move markets? *Review of Financial Studies, 22*(1), 151–186.

Barber, B. M., Odean, T., & Zhu, N. (2009b). Systematic noise. *Journal of Financial Markets, 12*(4), 547–569.

Bing, L., Chan, K. C., & Ou, C. (2014). Public sentiment analysis in twitter data for prediction of a company's stock price movements. In e-business engineering (ICEBE), 2014 I.E. 11th International Conference on (pp. 232-239). IEEE.

Black, F. (1986). Noise. *The Journal of Finance, 41*(3), 528–543.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science, 2*(1), 1–8.

Brown, G. W., & Cliff, M. T. (2004). Investor sentiment and the near-term stock market. *Journal of Empirical Finance, 11*(1), 1–27.

Carlin, B. I., Longstaff, F. A., & Matoba, K. (2014). Disagreement and asset prices. *Journal of Financial Economics, 114*(2), 226–238.

Chordia, T., Roll, R., & Subrahmanyam, A. (2002). Order imbalance, liquidity, and market returns. *Journal of Financial Economics, 65*(1), 111–130.

Corsi, F. (2005). Measuring and modelling realized volatility: From tick-by-tick to long memory (Doctoral dissertation, University of Lugano).

Da, Z., Engelberg, J., & Gao, P. (2015). The sum of all FEARS investor sentiment and asset prices. *Review of Financial Studies, 28*(1), 1–32.

Das, S. R., & Chen, M. Y. (2007). Yahoo! For Amazon: Sentiment extraction from small talk on the web. *Management Science, 53*(9), 1375–1388.

De Long, J. B., & Shleifer, A. (1991). The stock market bubble of 1929: evidence from closed-end mutual funds. *The Journal of Economic History, 51*(03), 675–700.

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy*, 703–738.

Diether, K. B., Malloy, C. J., & Scherbina, A. (2002). Differences of opinion and the cross section of stock returns. *Journal of Finance*, 2113–2141.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in neural information processing systems, 9*, 155–161.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., & Sugimoto, C. R. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology, 66*(8), 1523–1545.

Fisher, K. L., & Statman, M. (2000). Cognitive biases in market forecasts. *The Journal of Portfolio Management, 27*(1), 72–81.

Gao, L., & Kling, G. (2008). Corporate governance and tunneling: Empirical evidence from China. *Pacific-Basin Finance Journal, 16*(5), 591–605.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(suppl 1), 5228–5235.

Gruca, T. S., Berg, J. E., & Cipriano, M. (2005). Consensus and differences of opinion in electronic prediction markets. *Electronic Markets, 15*(1), 13–22.

Hornik, K., & Grün, B. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1–30.

Hsiao, C. (2014). Analysis of panel data, 3rd edn. Econometric Society monographs 54. Cambridge University Press.

Keynes, J. M. (1936). *The general theory of employment, interest and money*. London: Macmillan.

Kim, K. J. (2003). Financial time series forecasting using support vector machines. *Neurocomputing, 55*(1), 307–319.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering, 30*(1), 25–36.

Lee, C., Shleifer, A., & Thaler, R. H. (1991). Investor sentiment and the closed-end fund puzzle. *The Journal of Finance, 46*(1), 75–109.

Mao, H., Counts, S., & Bollen, J. (2011). Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data. ArXiv E-prints, p. 10. Available from: http://arxiv.org/abs/1112.1051.

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The management Revolution: Exploiting vast new flows of information can radically improve your company's performance. But first you'll have to change your decision making culture'[2012] Harvard Business Review.

McGraw Hill Financial (n.d.). Dow Jones Averages | About the Averages | Overview. Retrieved August 12, 2015, from https://www.djaverages.com/?go=about-overview

Miller, E. M. (1977). Risk, uncertainty, and divergence of opinion. *The Journal of Finance, 32*(4), 1151–1168.

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement, 7*(2), 191–205.

Nash, M. S. (2001). Handbook of parametric and nonparametric statistical procedures. *Technometrics, 43*(3), 374–374.

Neal, R., & Wheatley, S. M. (1998). Do measures of investor sentiment predict returns? *Journal of Financial and Quantitative Analysis, 33*(04), 523–547.

Oliveira, N., Cortez, P., & Areal, N. (2013a). *On the predictability of stock market behavior using stocktwits sentiment and posting volume*, *In Progress in Artificial Intelligence (pp. 355–365)*. Berlin Heidelberg: Springer.

Oliveira, N., Cortez, P., & Areal, N. (2013b). Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from twitter. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics* (p. 31). ACM.

Oliveira, N., Cortez, P., & Areal, N. (2014, July). Automatic creation of stock market lexicons for sentiment analysis using StockTwits data. In *Proceedings of the 18th International Database Engineering & Applications Symposium* (pp. 115-123). ACM.

Pedersen T, Banerjee S (2011) WordNet::Stem, Retrieved August 05, 2015, from http://search.cpan.org/~tpederse/WordNet-Similarity-2.05/lib/WordNet/stem.pm.

Poteshman, A. M. (2001). Underreaction, overreaction, and increasing misreaction to information in the options market. *The Journal of Finance, 56*(3), 851–876.

Qian, X. (2014). Small investor sentiment, differences of opinion and stock overvaluation. *Journal of Financial Markets, 19*, 219–246.

Rao, T., & Srivastava, S. (2012, August). Analyzing stock market movements using twitter sentiment analysis. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining* (ASONAM 2012) (pp. 119-123). IEEE computer society.

Schmeling, M. (2009). Investor sentiment and stock returns: Some international evidence. *Journal of Empirical Finance, 16*(3), 394–408.

Schwert, G. W. (1998). Stock market volatility: Ten years after the crash (no. w6381). National Bureau of economic research.

Shiller, R. J. (2000). Measuring bubble expectations and investor confidence. *The Journal of Psychology and Financial Markets, 1*(1), 49–60.

Stoffman, N. S. (2008). Individual and institutional investor behavior. ProQuest.

Tay, F. E., & Cao, L. J. (2002). Modified support vector machines in financial time series forecasting. *Neurocomputing, 48*(1), 847–861.

Taylor, S. J., & Xu, X. (1997). The incremental volatility information in one million foreign exchange quotations. *Journal of Empirical Finance, 4*(4), 317–340.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance, 62*(3), 1139–1168.

Tetlock, P. C., Saar-tsechansky, M. A. Y. T. A. L., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance, 63*(3), 1437–1467.

Theil, H., & Nagar, A. L. (1961). Testing the independence of regression disturbances. *Journal of the American Statistical Association, 56*(296), 793–806.

Varian, H. R. (1985). Divergence of opinion in complete markets: A note. *The Journal of Finance, 40*(1), 309–317.

Verma, R., & Soydemir, G. (2009). The impact of individual and institutional investor sentiment on the market price of risk. *The Quarterly Review of Economics and Finance, 49*(3), 1129–1145.

Verma, R., & Verma, P. (2007). Noise trading and stock market volatility. *Journal of Multinational Financial Management, 17*(3), 231–243.

Verma, R., & Verma, P. (2008). Are survey forecasts of individual and institutional investor sentiments rational? *International Review of Financial Analysis, 17*(5), 1139–1155.

Wang, Y. H., Keswani, A., & Taylor, S. J. (2006). The relationships between sentiment, returns and volatility. *International Journal of Forecasting, 22*(1), 109–123.

Wei, Q., & Dunbrack Jr., R. L. (2013). The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PloS one, 8*(7), e67863.

Zhang, M., Jansen, B. J., & Chowdhury, A. (2011). Business engagement on twitter: A path analysis. *Electronic Markets, 21*(3), 161–175.