


Do stock bulletin board systems (BBS) contain useful information? A viewpoint of interaction between BBS quality and predicting ability

Xiong Xiong^{a,b}, Chunchun Luo^a, Ye Zhang^a, Shen Lin^{b,c} 

^a*College of Management and Economics, Tianjin University, Tianjin, China*

^b*China Center for Social Computing and Analytics, Tianjin University, Tianjin, China*

^c*PBC School of Finance, Tsinghua University, Beijing, China*

Abstract

This study explores whether information on internet stock bulletin board systems (BBS) is valuable for stock return prediction, taking advantage of data derived from the biggest stock BBS in China. Using a text classification algorithm, we find the online messages significantly predict stock return with negligible R -squared. However, we find that accuracy of individual BBS posts is below 50 percent and there is no distinction at prediction accuracy between high- and low-quality stock BBS. Due to the autocorrelation of stock returns, we argue that BBS predicts stock returns because of its reflection on the simultaneous stock return rather than revelation on valuable information.

Key words: Stock online information; Information quality; Searching; Return prediction

JEL classification: G14, G40

doi: 10.1111/acfi.12448

The authors would like to thank anonymous referees, seminar participants in Tianjin

1. Introduction and literature review

1.1. Introduction

One of the most intriguing sources of unofficial and qualitative information is the vast amount of user-generated content online. Scholars and practitioners alike increasingly call attention to the popularity of online investment forums among investors and other financial professionals (Antweiler and Frank, 2004). Online message boards, mostly based on bulletin board systems (BBS), have recently been at the forefront of this development. The internet stock message boards serve as an excellent tool for investors to obtain stock information and exchange their opinions easily and almost freely. A questionnaire by Shenzhen Stock Exchange, one of the two major stock exchanges in China, revealed that about 35 percent of investors have access to stock BBS to acquire relevant market information. Online message boards allow users to publish messages without any limitation on number of characters. Many of these messages are dedicated to discussion of public companies and trading ideas. The impact of the internet on the financial industry and financial markets is enormous. On the one hand, online message boards dramatically optimize the way that investors acquire information, communicate and initiate trades (Barber and Odean, 2001; Clemons and Hitt, 2001; Litan and Rivlin, 2001). On the other hand, however, online message boards are flushed with noise (Barber and Odean, 2001; Clemons and Hitt, 2001; Depken and Zhang, 2010).

As yet, there is a little research examining whether and how BBS messages are related to financial indicators. For instance, Antweiler and Frank (2004) studied the effect of messages posted on Yahoo! Finance and Raging Bull. Their study shows that the effect of messages on stock returns is statistically significant but economically small. A few recent studies have made a further step in exploring the quality of information (Fan *et al.*, 2005; Zhang, 2009). These studies mainly focus on the relationship between quality of information and poster's reputation. However, while these studies provide a first indication of the relationship between online sentiment (or information) and financial indicators, they are still limited in a number of ways.

Firstly, the limited sample data length limits the use of internet stock BBS information in asset pricing. Different from Baker Wurgler's sentiment indicator, internet stock BBS data only appeared less than 30 years ago. And in these 30 years, its rapid development and change in network products and properties make the samples in different periods have distinct characteristics. This factor is mixed with complex internet data noise, confusing researchers about which of the following elements is the key difficulty for investors to extract economically effective information from internet data: whether the data itself do not contain valid information, or the limitations of data mining tools restrict information mining. Statistically significant 'predictive power', however, has prompted such research to continue. Secondly, whether internet data

are emotion-driven, information-driven, or just noise-driven has always been the most controversial topic. From Antweiler and Frank (information) to more

evidence that stock BBS information with better quality of posts could better predict future returns. This indicates that the receivers of stock BBS information cannot find effective information for their stock trading decisions. Finally, cross-sectional regression also indicates no evidence that the quality of internet stock BBS will affect its predictive ability. Therefore, we believe the internet stock BBS may only be a place for investors to vent their emotions, which are affected by the returns during the same period, so it happens to significantly predict the future returns statistically. To prove that unofficial internet data contain effective information, researchers need more rigorous empirical results to support this conclusion, rather than relying on simple statistical significance only.

1.2. Literature review and research questions

Numerous studies have manifested previously that investors' sentiment derived from the content of news media, social networking platforms and search engines is one of the key factors influencing the stock market as well as the lagged effect of returns. Wysocki (1998) first studied the information on internet stock message boards that had an influence on stock markets. Using a sample of over 3,000 stocks listed on Yahoo! message boards, he found message-posting volume predicts changes in next-day stock trading volume and returns. Antweiler and Frank (2004) extracted the sentiment of 1.5 million messages from the stock-linked internet message board, Yahoo! Finance and Raging Bull, built the bullishness of messages based on Naïve Bayes algorithm classification to study the relationship between the bullishness of and the corresponding stock performance. The study proved that the impact of bullishness on stock returns was statistically significant but with limited economic contribution. Tetlock (2007) tested the interactions between the market-wide media and stock market returns using daily contents from the popular *Wall Street Journal* column called 'Abreast of the Market', which reported news regarding yesterday's market conditions and other related issues. Tetlock found that high media pessimism predicted downward pressure on market prices followed by a reversion to fundamentals. Similar results can be found in the Twitter data (Bollen *et al.*, 2011), Google searching data (Joseph *et al.*, 2011; Da *et al.*, 2014) and other proxy of online sentiment (Siganos *et al.*, 2014). On the other hand, first-hand information or short-term sentiment plays a crucial role in the price dynamics of developing markets such as the Chinese stock market. Zhu *et al.* (2017) found that firm-level media reports affect the probability of stock price crash in China. Qian *et al.* (2018) investigated the price efficiency during the 2015 Chinese stock market crash and found a less serious price delay after the crash indicating that negative information travels slowly only when investors are overconfident. Other types of information hidden in the marginal trading (Li *et al.*, 2018) and market microstructure (Chen *et al.*, 2017; Xiong *et al.*, 2017; Lv and Wuarket',

also well studied in the literature and all of them are regarded as significant factors in the market dynamics or short-term movement of asset prices.

However, some studies provided a limited explanation of investors' sentiment to predict stock returns. Tumarkin and Whitelaw (2001) found that message board activity did not predict industry-adjusted returns or abnormal trading volume, which is consistent with market efficiency. Yin and Tan (2017) found that mass media cannot select analysts with high forecast accuracy, which then misleads investors. Kim and Kim (2014) picked up more than 30 million posts of 91 companies on Yahoo! Finance during a 6-year period (from January 2005 to December 2010) in order to test the impact of investors' sentiment on stock returns, volatility as well as volume, and claimed that neither for the entire industry nor an individual company did investors' sentiment impact on prediction of future earnings. Instead, they noticed that investors' sentiment was influenced positively by previous stock price. Besides the variables of investors' sentiment derived from news media, social platforms and search engines, there are two types of sentiment proxy variables. One is from the stock market, which includes: (i) share turnover on the New York Stock Exchange, the number and average first-day returns on initial public offerings (IPOs), the equity shares in new issues and the dividend premium, and combinations of the closed-end fund discount (Baker and Wurgler, 2006, 2007; Huang *et al.*, 2016); (ii) net mutual fund redemptions (Neal and Wheatley, 1998); (iii) fluctuations in discounts of closed-end funds (Lee *et al.*, 1991); (iv) bid-ask spreads/turnover (Baker and Stein, 2004); and (v) the portfolio allocations to equity versus cash and fixed-income securities (Edelen *et al.*, 2010). The other is formed by survey and investigation, which includes: (i) CCI (Consumer Confidence Index) published by the Conference Board (CBIND) and the Survey Research Center of the University of Michigan (Lemmon and Portniaguina, 2006); (ii) investors' intelligence published by the American Association of Individual Investors (Solt and Statman, 1988; Fisher and Statman, 2000; Lee *et al.*, 2002; Brown and Cliff, 2004, 2005); and (iii) investors' sentiment on animusX (Lux, 2011).

Investors' sentiment is of great importance to stock markets because plenty of scholars are trying to characterize investors' sentiment from various perspectives to explore its influence on stock markets (Xu and Zhou, 2018). Nevertheless, whether investors' sentiment has a significant impact on foreign stock markets is controversial based on the literature above. Thus, the first issue to focus on is to examine the interaction between investors' sentiment and stock returns on the Chinese stock market. By means of the stock BBS of Oriental Wealth, the most-visited online stock BBS on China's stock markets, we need to explore whether the emerging and updating of online information in China's stock BBS are similar to foreign social platforms. In line with the efficient market hypothesis, the price contains all revealed information on this stock. Even though some new information comes from the stock message board, the price would reflect the new information without any substantial influence. But according to the literature above, numerous scholars argue that

investors' sentiment has positive prediction power on stock returns. Whether the sentiment tracked from the stock message board in China could also predict positively on stock returns becomes the first problem we need to solve. We would expect that:

H1: Stocks with high BBS bullishness in the past obtain higher returns than their low-bullishness counterparts in the short term.

Investors use the stock BBS as an important means for two-way communication like a social platform. Someone posts a message on the stock forum and others would post replies (agree or disagree), which could make the original poster pay attention to the content. In existing studies, some scholars started to focus on the quality of information and tried to figure out proxy variables to measure the quality of the posts, aiming to explore the post spreading mechanism and posters' reputation on the basis of post quality. Fan *et al.* (2005) proposed a theoretical model to describe network feedback loops that provide a continuous incentive to users' real self-expression. Gu *et al.* (2008) thought the reduction of noise in stock BBS and the posts of high quality could attract more users because of the increasing cost of information processing and information overload. To investigate posters' reputation, Zhang (2009) chose the posts on the online bulletin board of TheLion, Wall Street Pit, and built indicators to reflect the quality of information categorized by sentiment threshold. One-day follow-up opinion on yesterday's stock returns can earn a higher reputation. Additionally, the poster's reputation depended on the quality of posting content rather than the quantity of posting. Sprenger *et al.* (2014) established a sentiment indicator and quality indicator by collecting and extracting the content on Twitter so as to investigate the diffusion of information there. The findings showed above-average investment advice could gain more retweets and approval. Instead, from the standpoint of the individual, high-quality information could not earn more retweets. In other fields related to information quality, most scholars mentioned that poster's reputation showed significant positive correlations with high-quality posting information (Konana *et al.*, 2000; Litan and Rivlin, 2001; Resnick and Zeckhauser, 2002; Houser and Wooders, 2006).

The studies mentioned above show investors have the potential to identify information quality. When new information becomes available, investors would search for high-quality information on stock BBS to acquire abnormal returns. If investors do dig out high-quality information on stock BBS, the high-quality information which has already appeared there has more positive predicting ability about stock returns. The intuition of this approach is similar to the discussion about market efficiency by Jensen (1968) who argued that the market should be efficient if the most professional participants (mutual fund managers) in the stock market cannot obtain positive excess returns. In a similar way, if the main readers cannot discover useful information on the stock

BBS or they never even try to, it is untenable to argue that stock BBS is valuable for investment decisions. Therefore, focusing on China's stock market, this paper proposes to test the predicting function of high-quality information on Chinese stock message boards to stock returns. Therefore, we expect:

H2: A high quality of stock BBS in the past improves the predicting ability of bullishness on future return by showing a bigger spread between portfolios with high and low bullishness.

The remainder of this paper is organized as follows. Section 2 describes the data source, classification of posts on the Online Message Board, and then introduces related explanatory variables. Section 3 presents the descriptive statistical analysis, portfolio analysis and Fama–Macbeth cross-sectional regression analysis of our two main hypotheses. Section 4 concludes.

2. Data set and methodology

2.1. Data set and sample selection of stock BBS

We chose the ‘Oriental Wealth’ stock message board (<http://guba.eastmoney.com/>) as our data source for stock BBS because it is the most-visited online stock BBS for China's stock markets. As illustrated in the introduction, investors obtain stock information and exchange their opinions easily and almost freely on this website and it deserves special attention. We study the 3.5-year period between 5 January 2011 and 30 June 2014, holding a total of 843 trading days, to deal with stable developments on the China financial markets and to avoid potentially distorting repercussions of the turbulence in 2015. We focus on the Shanghai-Shenzhen 300 Index to adequately reflect the entire spectrum of China equities, including a wide range of industries. We limit our study to 286 companies from the stock components of the Shanghai-Shenzhen 300 Index, as 14 stocks are eliminated because of missing data or long trading suspensions. It gives priority to the crawling data on stock BBS related to text messages of posting titles and daily postings of each stock because posting titles show posters' main point of view to the future trend of the stock market, and the number of daily postings describes the activity level of each stock.

The stock data are obtained from the CSMAR database, the biggest financial database in China and the only database which is included in Wharton Research Data Services (WRDS) as stock trading data for the stock market in China. The data obtained include: daily return, trading volume, turnover, book-to-market ratio, market equity, market return, *Small-Minus-Big* return (size factor) and *High-Minus-Low* return (value factor).

2.2. Sentiment classification

In order to examine the relationship between signals from stock BBS and market movements, we had to classify messages into three types: positive, neutral and negative. As our data set contains too many messages for manual classification, we chose to classify messages automatically using well-established methods from computational linguistics.

Chinese sentiment classification consists of two steps: word segmentation and sentiment classification. Word segmentation is usually unnecessary in English sentiment classification because the words in an English sentence are separated naturally. However, Chinese sentences are composed of Chinese characters. Characters usually have their own meaning, such as ‘好’ means good. But generally, one word is composed of 2–4 characters, and the meaning of the characters is not always the same as the word. For example, the word ‘多头’ means bull side of the market, but neither the character ‘多’ nor ‘头’ has the same meaning. Therefore, we must divide the sentences into the words appropriately and the meaning must remain the same. We employ the ‘FudanNLP-1.6.1’ software as our word segmentation instrument, which is also widely used in other studies of natural language processing for Chinese text (Li *et al.*, 2015).

The process of sentiment classification applied in this paper is same as in Antweiler and Frank (2004): Naïve Bayesian Classification (NBC). It is a basic but functional classification in human natural language. We provide the technical details in the Appendix and only report the classification results in Table 1.

We use 5,000 training samples with NBC to determine the in-sample accuracy. In Antweiler and Frank (2004), the same NBC was used to classify English messages and their in-sample accuracy was 88.1 percent with 1,000 manual samples. Our result is given in Panel A of Table 1. Our in-sample accuracy is 85.4 percent, which is a little lower than the accuracy in Antweiler and Frank (2004). One possible reason is that it is difficult to recognise tone and sarcasm in Chinese. Moreover, we randomly chose and manually classified a further 1,000 messages for our out-sample test. Panel B of Table 1 shows the accuracy of the out-sample classification. The accuracy declines from 85.4 to 77.9 percent, which is also at a high level. We cannot compare our out-sample accuracy with Antweiler and Frank (2004) because they do not report it. However, our accuracy is higher than other studies with English classification (e.g. Das and Chen, 2007; Kim and Kim, 2014). Most importantly, the situation of messages has been classified to opposite sentiment (the messages classified as positive ones by manual but negative ones by NBC and vice versa) keep a low percentage (0.4 and 0.2 percent). This result ensures our classification does not have systematic error.

2.3. Bullishness of stock BBS

This paper is aimed at determining the predictive effect of information in stock BBS on stock returns. To test the relationship between the fluctuation of stock returns and thousands of daily messages in stock BBS, we need to extract

Table 1
Accuracy of Naïve Bayesian classification

		Naïve Bayesian learning classification		
Manually classify	%	Positive	Neutral	Negative
Panel A. In-sample classification accuracy				
Positive	15.6	10.9	4.4	0.3
Neutral	62.6	1.6	58.8	2.2
Negative	21.8	0.1	6.0	15.7
5,000 messages		12.6	69.2	18.2
Panel B. Out-sample classification accuracy				
Positive	12.5	7.1	5.0	0.4
Neutral	69.3	4.9	60.5	3.9
Negative	18.2	0.2	7.7	10.3
1,000 messages		12.2	73.2	14.6

The titles of 6,000 posts are randomly picked as our trading data set of which 5,000 are used for Naïve Bayesian conditional probability trading and the other 1,000 posts are used for the out-sample test. We manually classified these 6,000 posts into three types, which are the labels for trading and testing. This table reports the probabilities of manually classification, NBC-based classification and the cross-probabilities of both classifications. Panels A and B show the in-sample (5,000 posts) and out-sample (1,000 posts) results, respectively.

messages to a firm-specific variable called Bullishness (referred to as *Bul*) in accordance with the definition of Antweiler and Frank (2004).

Based on the classified messages and the bullishness building method of Antweiler and Frank (2004), we define buying as 1, holding as 0, and selling as −1. M_t^{buy} (M_t^{sell}) represents the number of buy (sell) signals on day t . We follow Antweiler and Frank (2004) by defining bullishness as:

$$Bul_t = \ln \left(\frac{1 + M_t^{buy}}{1 + M_t^{sell}} \right). \quad (1)$$

The bullishness index considers not only the changes of signals but also the number of messages giving greater weight to a more robust larger number of messages expressing a particular opinion. A high value for Bullishness means more postings which support buying. The more investors who claim positive opinions of the trend in the stock market, the greater possibility that they believe a rise in stock price will occur.

2.4. Quality index of stock BBS

Quality of stock BBS should precisely estimate the standpoints of the trend of the stock market to one stock from all posters and become a signal for other

investors to evaluate the quality of online message of one stock in BBS. According to the study by Zhang (2009), who studied the determinants of poster reputation on online message boards, we define the quality as forecast accuracy rate of postings which involves the investing opinions about stock i on day $t-1$ to the return of stock i on day t . The value range of Quality is 0–1, where 0 means the forecast accuracy of stock i on day $t-1$ is 0 percent and 1 means the forecast accuracy of stock i on day $t-1$ is 100 percent. We follow Zhang (2009) by defining Quality as:

$$Q_{i,t} = \frac{\sum_{j=1}^n q_{i,t}^j}{n}, \quad (2)$$

where n is the total number of postings of stock i on day $t-1$ in stock BBS and q is a dummy variable, which is equal to 1 when sentiment direction of post j of stock i on day $t-1$ is the same as the return of stock i on day t . Otherwise, $q_{i,t}^j = 0$. The formula is as follows:

$$q_{i,t}^j = \begin{cases} 1, & \text{if } S_{i,t-1}^j R_t^i > 0 \\ 0, & \text{otherwise} \end{cases}, \quad (3)$$

where $S_{i,t-1}^j$ represents the opinion of post j of stock i at day $t-1$ ($S_{i,t-1}^j \in \{-1, 1\}$), excluding the posts with ‘opinion of holding (0)’ when counting quality of stock BBS. We do not consider the post with $S_{i,t-1}^j = 0$ due to meaningless posts, such as one-character posts, whose number is highly volatile. These meaningless posts may significantly dilute the quality measurement if they are considered. There is a small difference between our measurement and Zhang’s where there is one-day lag between post and return when we define the dummy $q_{i,t}^j$, whereas Zhang use the simultaneous return to evaluate the correctness of posts. The reason is that the quality of BBS posts should be measured by their predicting ability on future returns rather than their reaction to the same period return. Therefore, if a post makes a correct prediction about future one-day returns, we consider it as a high-quality post with $q_{i,t}^j = 1$.

Figure 1 gives the fluctuation of overall quality of stock BBS posts between 5 January 2011 and 30 June 2014. We can see that the quality of stock BBS fluctuates around 0.45. The mean quality of posts with 286 stocks during the period is 0.436, which is below 0.5, illustrating that the prediction capacity of an individual post is worse than random predictions.

In the following analysis, we try to identify the posts that carry a high reputation in the past whose Q_t is high. We use a dummy variable Qd_t , which equals 1 when Q_t is higher than the threshold value 0.5, to define a high-quality stock forum. Therefore, the bullishness on day t interacted with $Qd_t = 1$ should give a higher prediction ability if investors are trying to find higher quality information.

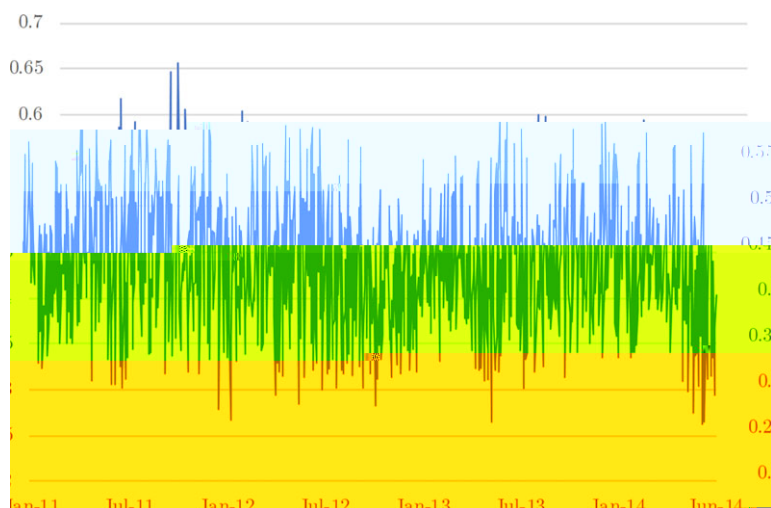


Figure 1 The fluctuation of overall quality of stock bulletin board systems. The x -axis represents the 843 trading days of our sample period. The y -axis marks the value of overall quality of given days.

Some other commonly used return predictors are also considered, including market equity (ME), book-to-market ratio (BM) and turnover rate ($Turnover$). Table 2 presents definitions of all the variables in this paper.

3. Empirical methods and main results

3.1. Summary statistics and single-sorted portfolios

This section reports summary statistics and the results for single-sorted portfolios. During our sample period, we have collected 8.8 million stock-related posts on Shanghai-Shenzhen 300 Index companies. The number of postings for individual stocks ranges from 0 to 5,600 daily, and all the postings

Table 2
Variable definitions

Ret_t	Daily return
Bul_t	Bullishness
MV_t	Number of daily posts
Q_t	Quality index of stock BBS
Qd_t	High-quality dummy variable, $Q_t > 0.5$, $Qd_t = 1$. Otherwise, $Qd_t = 0$
ME_t	Market equity
BM_t	Book-to-market ratio
$Turnover_t$	Turnover rate

represent an average of 10,600 postings per trading day for all listed stocks. An average 36.5 posts per company-day indicates that our data set comprises a dense information stream.

Table 3 presents summary statistics of stocks' BBS and firm characters sorted on bullishness (Panel A) and quality (Panel B). At the end of each day t , we sort stocks into quintiles based on bullishness (Panel A) or sort stocks into three portfolios based on quality (Panel B). In each group, the average stock number, the ex-ante portfolio return (Ret_t), is calculated as average return of stocks in each portfolio on day t . Bul_t is the equal-weighted bullishness of stocks in each portfolio on day t . Q_t is the average quality of stocks' BBS in each portfolio on day t . $Bul_t * Q_d$ is the average value of the interaction term of lagged high-quality dummy and bullishness. We also calculate other firm characteristics such as nature logarithm of market equity ($\ln(ME_t)$), book-to-market ratio (BM_t), accumulated return in the past five trading days ($Ret_{t-5,-1}$) and share turnover ($Turnover_t$) for each quintile in the same way.

Panel A of Table 3 reports summary statistics for portfolios sorted by bullishness. High-bullishness firms tend to have lower book-to-market ratios, higher past returns and higher turnover rate than low-bullishness firms. In particular, stocks with high-bullishness outperform stocks with low-bullishness during the same period which means the sentiment released in the BBS may be caused by same-day returns. The return spread between top and bottom quintiles is 102 basis-points in equal-weighted and 78 basis-points in value-weighted per day. In addition, the spread between quality of stock BBS for the high- and low-bullishness portfolios is -3 basis points (BPs). High-bullishness firms tend to have lower quality of stock BBS, which indicates that investors are more susceptible to negative sentiment than positive sentiment. Bullishness with high past quality shows the same trend as bullishness. But the spread between the bullishness of high quality for the high- and low-bullishness is 10 BPs per day, which is much slighter compared with the spread between the highest bullishness and the lowest bullishness. Three of the five portfolios have negative average bullishness and the average of bullishness of the fourth one is close to 0, which indicates that most investors lean more towards bearish sentiment on the stock market in stock BBS of 'Oriental Wealth' during the selected time period.

Panel B in Table 3 presents summary statistics for portfolios sorted by quality of stock BBS. High-quality of stock BBS firms tend to have lower market equity, lower book-to-market ratios and higher turnover rate than low-quality of stock BBS firms. In particular, stocks with high-quality of stock BBS underperform stocks with low-quality of stock BBS, due to the lower quality of stock BBS tending to have higher bullishness. Due to the overall trend of the stock market during our sample period being slightly downward, a more negative bullishness tends to have a higher quality, while the simultaneous return could be lower as shown in Panel A. The spread between top and bottom portfolios is -39 , -7 and -0.6 BPs for return, bullishness and bullishness of high quality, respectively.

Table 3
Simultaneous return and related features of portfolio

	<i>N</i>	<i>Ret_t</i> (<i>EW</i>)	<i>Ret_t</i> (<i>VW</i>)	<i>Bul_t</i>	<i>Q_t</i>	(<i>Bul_t</i> * <i>Q_{d,t}</i>)	ln(<i>ME_t</i>)	<i>BM_t</i>	<i>Ret_{t-5,-1}</i>	<i>Turnover_t</i>
Panel A: Portfolio of bullishness										
<i>Bul_L</i>	57	−0.51	−0.36	−1.40	0.49	−0.20	23.56	0.60	−0.67	0.77
<i>Bul₂</i>	57	−0.28	−0.21	−0.75	0.47	−0.17	23.61	0.60	−0.16	0.81
<i>Bul₃</i>	57	−0.05	−0.06	−0.32	0.46	−0.15	23.72	0.60	−0.76	0.87
<i>Bul₄</i>	57	0.16	0.15	0.03	0.44	−0.12	23.47	0.59	1.52	0.80
<i>Bul_H</i>	58	0.51	0.41	0.72	0.45	−0.10	23.57	0.58	0.63	0.83
<i>Bul_{IR-Bul_L}</i>	–	1.02 ^c	0.78 ^c	2.12 ^c	−0.03 ^c	0.10 ^c	0.01	−0.01 ^c	1.30	0.05 ^c
<i>t</i> -stat	–	(57.6)	(32.2)	(548)	(−11.2)	(26.6)	(1.6)	(−4.8)	(1.3)	(7.5)
Panel B: Portfolio of quality										
<i>Q_L</i>	95	0.15	0.17	−0.31	0.11	−0.15	23.57	0.60	−0.74	0.74
<i>Q_M</i>	95	−0.01	0.03	−0.32	0.47	−0.14	23.66	0.58	1.45	0.87
<i>Q_H</i>	96	−0.23	−0.29	−0.38	0.80	−0.15	23.53	0.59	−0.36	0.84
<i>Q_H-Q_L</i>	–	−0.39 ^c	−0.46 ^c	−0.07 ^c	0.68 ^c	−0.006 ^b	−0.03 ^c	−0.01 ^c	0.37	0.10 ^c
<i>t</i> -stat	–	(−18.8)	(−20.3)	(−10.7)	(292)	(−2.0)	(−4.4)	(−4.6)	(0.5)	(15.0)

At the end of each day *t*, all sample stocks are grouped into five portfolios based on their bullishness index and BBS quality calculated based on the daily posts in the stock BBS. The ex-ante equal-weighted stock characters are given in the table, including portfolio return of day *t* in both equal-weighted and value-weighted, bullishness index, BBS quality, cross-term of bullishness and quality dummy, nature-logarithmic market equity, book-to-market ratio, return during last 5 days skipping 1 day and share turnovers in day *t*. Panel A shows the results of bullishness portfolios and Panel B shows the results for BBS quality. The superscripts a, b, c represent the 90, 95 and 99 percent significance levels, respectively.

Table 4
 Portfolios' returns at day $t + 1$ sorted by bullishness

Bul_L	Bul_2	Bul_3	Bul_4	Bul_H	Bul_H-Bul_L

3.2. Bullishness index and stock return

3.2.1. Portfolio analysis

To describe the relationship between bullishness and returns, we keep the classification of portfolios in Section 3.1 and calculate the portfolio returns at day $t + 1$ from January 2011 to June 2014. According to the summary statistics in Section 3.1, portfolios categorized by various factors have correlations to

size and book-to-market. To remove the influence of corporate characteristics, we use the Fama–French three-factors model to deduce the intercept term of different portfolios and depict the relationship between bullishness and returns in detail.

Table 4 shows the equal-weighted return of day $t + 1$, corresponding alphas based on the Fama–French three-factors model and value-weighted returns of each portfolio. They are given in the Panels A, B and C, respectively. Figure 2 is a graphical representation of the equal-weighted result of Table 4. As shown in Figure 2, higher bullishness tends to have a higher future return. The difference between Bul_H and Bul_L is 11 BPs, which is not negligible if it is a realized profit as a daily return spread. The alpha of the Fama–French three-factor model and value-weighted return show similar results to the equal-weighted one, which confirm the robustness of the difference between Bul_H and Bul_L . However, due to the transaction fee and bid–ask spread, its high transaction frequency vanishes any economic meaning of this tiny return. However, we still cannot deny that bullishness has a positive impact on returns and systemic risk exposure cannot explain this phenomenon, even though the predicting ability is economically insignificant.

3.2.2. Fama–Macbeth cross-sectional regressions

The sorting result shown in the Section 3.2 is simple and intuitive, but it cannot explicitly control for other variables that may influence future returns. Multiple sorting can solve this problem but sorting on three or more variables is impractical. Thus, to examine other possible mechanisms, we perform Fama–MacBeth cross-sectional regressions (Fama and MacBeth, 1973), which allows us to control for additional variables conveniently on the previous research question.

The most intuitive idea about the result shown above is that valuable information is hidden in the enormous BBS data and that investors fail to reflect it immediately into the trading price. However, in the summary statistics, we also show that the BBS sentiment or information highly correlated with the return in the same period. The auto-correlation of stock return is a well-known stylized fact. It could be an alternative explanation to the predictive ability of the stock BBS. Therefore, to conclude the confirmation of our first hypothesis, we need to look deeply into the substitutional relation between stock BBS and past return.

Antweiler and Frank (2004) adopted a panel regression method which tests the time-series and cross-sectional effect of bullishness simultaneously. However, it is difficult to identify which component dominates the pricing effect. To solve the existing problems, we focus on the cross-sectional effect and choose Fama–MacBeth regression to check the relationship between bullishness of stock BBS and corresponding stock return.

As we mentioned above, the most concerning problem is that bullishness is associated with past stock market performance, such as past returns or trading

Table 5
Cross-sectional predicting ability of bullishness

	R_{t+1}			
	(1)	(2)	(3)	(4)
Bul_t	0.055 ^c (9.50)	0.051 ^c (9.35)	0.035 ^c (6.53)	0.037 ^c (6.72)
$\ln(MV_t)$		−0.015 ^c (−2.60)	−0.001 (−0.22)	−0.000 (−0.01)
R_t			0.040 ^c (6.59)	0.038 ^c (6.21)
R_{t-1}				−0.007 (−1.30)
$Turnover_t$			−0.001 ^c (−4.16)	−0.001 ^c (−4.17)
Ave. R -squared	0.006	0.016	0.055	0.078
Number of obs.	841	841	841	841

Each day, we run a cross-sectional regression of returns on lagged variables including the bullishness of stock BBS and other possible return predictors. The definitions of variables are given in Table 2. This table reports the time-series average of the coefficients and R -squared. The t -statistics based on Newey–West standard errors with 20 lags are given in parentheses. The superscripts a, b, c represent the 90, 95 and 99 percent significance levels, respectively.

volume. We examine the impact of bullishness on returns and add the control variables one by one to rule out the possible past trading information carried by our main research variable. We control several traditional return predictors, such as past returns, past trading volume and number of BBS posts. We do not control the industry effect due to the limitation of our sample firms which is only 287. Also, it is hardly convincing that BBS bullishness is highly correlated with industry. The result is shown in Table 5. Considering that investor attention, measured by number of posts, is one of the key factors illustrated in BBS, Regression (2) in Table 5 added the number of posts as a control variable at first. Regression (3) added return and turnover with 1-day lag as control variables due to the strong first-order autocorrelation of returns in China and the attention effect included in BBS bullishness (Antweiler and Frank, 2004). Regression (4) added return with 2-day lag as control variable to control the second-order autocorrelation of returns.

Table 5 proves that Hypothesis 1 is tenable but the information argument is dubious based on the results of the Fama–MacBeth regression. The result in Regression (1) shows that the coefficient of bullishness is significant and positive ($\beta = 0.055$, t -stat = 9.50). Regardless of whether adding control variables or not, bullishness remains significantly positive and this is a confirmation of the result of our previous portfolio analysis. However, even though the coefficient is significantly positive, the R -squared of univariate

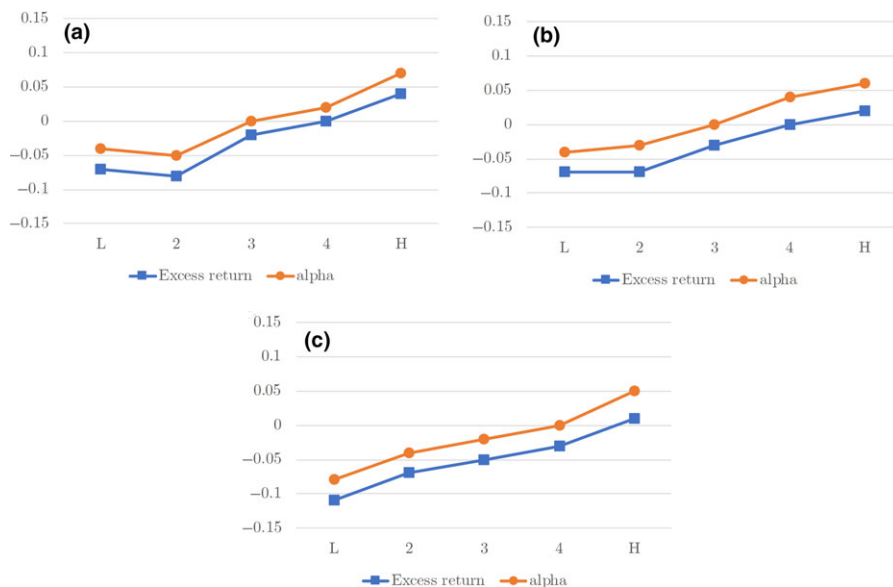


Figure 3 Portfolio returns of different bullishness stocks with low (panel a), median (panel b) and high (panel c) quality. The y-axis marks the average daily equal-weighted return of each portfolio in percent.

3.3. High quality stock BBS and predictive ability of bullishness

3.3.1. Portfolio analysis

In this section, we use a double-sorting approach to investigate the relationship between information quality and the predictive ability of bullishness in stock BBS. We divide all sample stocks into three portfolios based on previous day BBS quality from low to high and then divide each of them into five sub-portfolios based on bullishness, respectively. We get a total of 15 portfolios with different past information quality and different bullishness signal. We calculate the equal- and value-weighted returns of portfolios at day $t + 1$ from January 2011 to June 2014 and work out the Fama–French three-factor alpha of each portfolio to exclude the impact of corporate characteristics. If we find different return spread of high-low bullishness portfolio between different quality stock BBS, it can be concluded that the information quality impacts the prediction power of stock BBS, implying there is available information hidden in them.

Panel A in Table 6 shows the average returns of portfolios sorted by information quality and bullishness, and Figure 3 is the graphical representation of Table 6. In Table 6, we observe that there is no significant difference in return spread of $Bul_H - Bul_L$ among different information quality. The spread is

future stock returns better, in which the underlying story is the stock BBS contains useful information for future returns. To explore whether investors could find the high-quality information for their trading decision making, we employ a similar approach to that used in Section 3.2.2.

According to our hypothesis, the cross term of past quality and bullishness should significantly positively predict future returns as well as that the prediction power of bullishness should decrease when investors are trying to find high-quality information in stock BBS. For this reason, we regress the stock return in $t + 1$ on lagged bullishness and the cross term of Bul_t and the indicator of high-quality stock BBS Qd_{t-1} . As with Section 3.2.2, we add the number of total posts, past return and volume as control variables.

Table 8
Robustness test

	R_{t+1}			
	(1)	(2)	(3)	(4)
Bul_t		0.041 ^c (8.92)	0.045 ^c (7.99)	0.042 ^c (7.97)
$Qd_t * Bul_t$			−0.009 (−1.11)	−0.013 (−1.55)
Qd_t			0.016 ^a (1.90)	0.017 ^a (1.92)
$\ln(MV_t)$				−0.001 (−0.02)
R_t	0.035 ^c (5.92)	0.030 ^c (5.22)	0.031 ^c (5.30)	0.030 ^c (5.12)
$R_{t-4,-1}$	−0.000 (−0.22)	−0.000 (−0.26)	−0.000 (−0.24)	−0.000 (−0.26)
$Turnover_t$	−0.047 ^c (−3.14)	−0.044 ^c (−3.58)	−0.044 ^c (−3.59)	−0.029 ^b (−2.09)
$Turnover_{t-4,-1}$	0.000 (0.82)	0.000 (0.82)	0.000 (0.91)	0.000 (0.80)
$\ln(ME)$	−0.013 ^a (−1.89)	−0.014 ^a (−1.92)	−0.013 ^a (−1.92)	−0.019 ^b (−2.16)
BM	0.020 (0.76)	0.026 (0.96)	0.025 (0.92)	0.036 (1.28)
Ave. R -squared	0.105	0.109	0.113	0.119
Number of obs.	841	841	841	841

Each day, we run a cross-sectional regression of returns on lagged variables including the bullishness of stock BBS and other possible return predictor. The definitions of variables are given in Table 2. This table reports the time-series average of the coefficients and R -squared. The t -statistics based on Newey–West standard errors with 20 lags are given in parentheses. The superscripts a, b, c represent the 90, 95 and 99 percent significance levels, respectively.

Table 9
Cross-sectional predicting ability of bullishness on long-run return

	$R_{t+1,t+5}$				
	(1)	(2)	(3)	(4)	(5)
Bul_t	0.060 ^c (2.63)	0.050 ^b (2.25)	0.057 ^c (2.88)	0.054 ^c (2.87)	0.065 ^c (3.30)
$Qd_t^*Bul_t$	−0.005 (−0.24)	−0.006 (−0.31)	−0.004 (−0.18)	0.009 (0.45)	
Qd_t	0.017 (0.67)	0.021 (0.87)	0.019 (0.85)	−0.004 (−0.18)	
$\ln(MV_t)$		−0.071 ^c (−2.73)	−0.030 (−1.05)	−0.024 (−0.81)	−0.022 (−0.77)
R_t			1.685 (0.95)	0.513 (0.30)	0.499 (0.30)
R_{t-1}				−2.023 ^a (−1.65)	−2.072 (−1.65)
$Turnover_t$			−0.009 (−1.56)	−0.009 (−1.51)	−0.009 (−1.52)
Ave. R -squared	0.015	0.024	0.073	0.085	0.077
Number of obs.	841	841	841	841	841

Each day, we run a cross-sectional regression of future 5-day returns on lagged variables including the bullishness of stock BBS and other possible return predictor. The definitions of variables are given in Table 2. This table reports the time-series average of the coefficients and R -squared. The t -statistics based on Newey–West standard errors with 20 lags are given in parentheses. The superscripts a, b, c represent the 90, 95 and 99 percent significance levels, respectively.

Table 7 describes the result of the Fama–MacBeth regression. The regression results show that bullishness keeps good predictive ability positively and significantly (coefficient = 0.062, t -statistics = 7.89) after adding the cross term whose regression coefficient is negative and non-significant (coefficient = −0.010, t -statistics = −1.09). The result of the regression with control variables also shows bullishness keeps its predicting ability positively and significantly. It confirms the result of portfolio analysis that high information quality contributes little or nothing to the return prediction from stock BBS bullishness. The BBS participants were not trying or failed to identify high-quality information from the massive amount of internet data whose information density is too low or even just simply noisy.

3.4. Robustness testing

Table 3 indicates that portfolio returns might be related to corporate characteristics such as book-to-market and market value. Our statement could

be incorrect if the impact of quality is correlated with these factors. To make sure of the robustness of our results and to rule out this possibility, we input the corporate characteristics as control variables into the cross-sectional regression. Two well-known characteristics, market equity and book-to-market ratio, from the Fama–French (Fama and French, 1993) three-factors model are considered as our main controls. We employ Fama–MacBeth cross-sectional regression only with control variables and then we add *Bul*, cross term of *Qd* and *Bul*, $\ln(MV)$ into the regression, respectively, to look deeply into their impact on stock returns.

Table 8 presents the results of robustness testing. The result of Regression (1) shows returns have significant first-order autocorrelation and the difference in daily returns can be explained by market value. After adding control variables such as corporate characteristics and number of posts, the results of regression equations (2)–(4) is still in accordance with the results in the previous section, which deny the information story of stock BBS. Bullishness still has significant and positive predicting ability but a very small contribution to *R*-squared. Cross term of quality and bullishness cannot predict returns, rejecting H_2 , which means investors cannot acquire high-quality information in stock BBS.

Another concern is that BBS information predicts the long-run future returns such as 5-day returns. To test its long-term impact on stock returns, we replace the dependent variable with 5-day return and similar regression as in Tables 5 and 7 are investigated. Results are given in Table 9 which shows that quality does not improve the predicting ability of BBS information. The contribution of bullishness to *R*-squared remains small, which confirm our main argument on long-term returns prediction.

We argue that, if the stock BBS contains useful information for investor trading, the BBS with high quality message should get more attention than the one with low quality, which should impact the predicting ability of BBS bullishness. However, we find no evidence that high quality BBS predicts future returns better. Combined with the tiny *R*-squared of the regression of Bullishness, we conclude that the messages in BBS do not contain useful information of stock for investors and its statistically significant predicting ability is inherited from reflection on the simultaneous stock returns.

4. Conclusions

As internet technologies have been widely developed and applied at a snowballing pace, numerous financial scholars have tried to distinguish useful information in internet data. There are also many scholars who believe that internet forums contain effective information that can reveal the fundamentals of the company. This paper employs data from the stock BBS of ‘Oriental Wealth’, the most-visited online stock BBS on the stock market in China. We indirectly study the possibility of effective information contained in stock BBS from another perspective: the relationship between information quality and prediction ability.

We designed two empirical tests. We first extract messages to a firm-specific variable called bullishness following the method of Antweiler and Frank (2004). If stock BBS contains useful information, bullishness should be able to effectively predict the future changes of stock returns. Using a cross-sectional method, we find that bullishness in China stock BBS statistically significantly predict the next-day stock returns with small coefficient and *R*-squared. Due to the autocorrelation of stock returns, we consider that stock BBS predicts the stock returns because of its reflection on the stock returns in the same period rather than revealing available information that is not priced in the stock market.

Secondly, we follow the method of Zhang (2009) for quality measurement of stock BBS. If stock BBS contains useful information, investors would search for high-quality information there to acquire abnormal returns. If investors do dig out high-quality stock BBS, the high-quality information that has already appeared in stock BBS should have greater predictive power on stock returns. However, we find no evidence that high-quality BBS predicts the future returns better. Neither portfolio analysis nor Fama–Macbeth cross-sectional regression shows that bullishness in a higher quality BBS better predicts or has a higher impact on future stock returns.

Based on these results, we conclude that the messages in BBS hardly contain useful information for future stock pricing and its predicting ability is inherited from the reflection on simultaneous stock returns and their autocorrelation. Perhaps there is a little useful information in the BBS, but the information density in BBS is remarkably low, destroying the possibility of discovery of useful information. These findings suggest that if researchers want to argue that the non-authoritative internet information is not released randomly by noise traders and contain useful information for stock trading, more robust evidence that is not only statistically but also economically significant, and a clear influence channel from this information to stock performance, should be provided.

References

- Antweiler, W., and M. Z. Frank, 2004, Is all that talk just noise? The information content of internet stock message boards, *The Journal of Finance* 59, 1259–1294.
- Baker, M., and J. C. Stein, 2004, Market liquidity as a sentiment indicator, *Journal of Financial Markets* 7, 271–299.
- Baker, M., and J. Wurgler, 2006, Investor sentiment and the cross-section of stock returns, *The Journal of Finance* 61, 1645–1680.
- Baker, M., and J. Wurgler, 2007, Investor sentiment in the stock market, *Journal of Economic Perspectives* 21, 129–152.
- Barber, B. M., and T. Odean, 2001, The internet and the investor, *Journal of Economic Perspectives* 15, 41–54.
- Bollen, J., H. Mao, and X. Zeng, 2011, Twitter mood predicts the stock market, *Journal of Computational Science* 2, 1–8.
- Brown, G. W., and M. T. Cliff, 2004, Investor sentiment and the near-term stock market, *Journal of Empirical Finance* 11, 1–27.

- Brown, G. W., and M. T. Cliff, 2005, Investor sentiment and asset valuation, *The Journal of Business* 78, 405–440.
- Chen, X., Y. Liu, and T. Zeng, 2017, Does the T + 1 rule really reduce speculation? Evidence from Chinese Stock Index ETF, *Accounting and Finance* 57, 1287–1313.
- Clemons, E. K., and L. Hitt, 2001, Financial services: transparency, differential pricing, and disintermediation, in: R. Litan, A. Rivlin, eds., *The Economic Payoff from the Internet Revolution* (Brookings Institution, Washington, DC), 87–128.
- Da, Z., J. Engelberg, and P. Gao, 2014, The sum of all FEARS investor sentiment and asset prices, *The Review of Financial Studies* 28, 1–32.
- Das, S. R., and M. Y. Chen, 2007, Yahoo! for Amazon: sentiment extraction from small talk on the web, *Management Science* 53, 1375–1388.
- Depken, C. A. II, and Y. Zhang, 2010, Adverse selection and reputation in a world of cheap talk, *The Quarterly Review of Economics and Finance* 50, 548–558.
- Dong, Z., and Q. Dong, 2003, HowNet – a hybrid language and knowledge resource, IEEE Proceedings of International Conference on Natural Language Processing and Knowledge Engineering (Los Alamitos), 820–824.
- Edelen, R. M., A. J. Marcus, and H. Tehranian, 2010, Relative sentiment and stock returns, *Financial Analysts Journal* 66, 20–32.
- Fama, E. F., and K. R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.
- Fama, E. F., and J. D. MacBeth, 1973, Risk, return, and equilibrium: empirical tests, *Journal of Political Economy* 81, 607–636.
- Fan, M., Y. Tan, and A. B. Whinston, 2005, Evaluation and design of online cooperative feedback mechanisms for reputation management, *IEEE Transactions on Knowledge and Data Engineering* 17, 244–254.
- Fisher, K. L., and M. Statman, 2000, Investor sentiment and stock returns, *Financial Analysts Journal* 56, 16–23.
- Gu, B., P. Konana, and H. W. M. Chen, 2008, Melting-pot or homophily? An empirical investigation of user interactions in virtual investment-related communities. Working paper (University of Texas).
- Houser, D., and J. Wooders, 2006, Reputation in auctions: theory, and evidence from eBay, *Journal of Economics and Management Strategy* 15, 353–369.
- Huang, H., G. Jin, and J. Chen, 2016, Investor sentiment, property nature and corporate investment efficiency: based on the mediation mechanism in credit financing, *China Finance Review International* 6, 56–76.
- Jensen, M. C., 1968, The performance of mutual funds in the period 1945–1964, *The Journal of Finance* 23, 389–416.
- Joseph, K., M. B. Wintoki, and Z. Zhang, 2011, Forecasting abnormal stock returns and trading volume using investor sentiment: evidence from online search, *International Journal of Forecasting* 27, 1116–1127.
- Kim, S. H., and D. Kim, 2014, Investor sentiment from internet message postings and the predictability of stock returns, *Journal of Economic Behavior and Organization* 107, 708–729.
- Konana, P., N. M. Menon, and S. Balasubramanian, 2000, The implications of online investing, *Communications of the ACM* 43, 34–41.
- Lee, C. M., A. Shleifer, and R. H. Thaler, 1991, Investor sentiment and the closed-end fund puzzle, *The Journal of Finance* 46, 75–109.
- Lee, W. Y., C. X. Jiang, and D. C. Indro, 2002, Stock market volatility, excess returns, and the role of investor sentiment, *Journal of Banking and Finance* 26, 2277–2299.
- Lemmon, M., and E. Portniaguina, 2006, Consumer confidence and asset prices: some empirical evidence, *The Review of Financial Studies* 19, 1499–1529.

- Li, H., J. Ding, D. Nie, and L. Tang, 2015, Accurate recommendation based on opinion mining, in: H. Sun, C. Y. Yang, C. W. Lin, J. S. Pan, V. Snasel, A. Abraham, eds., *Advances in Intelligent Systems and Computing*, vol. 329 (Springer, Cham, Switzerland), 399–408.
- Li, R., N. Li, J. Li, and C. Wu, 2018, Short selling, margin buying and stock return in China market, *Accounting and Finance* 58, 477–501.
- Litan, R. E., and A. M. Rivlin, 2001, Projecting the economic impact of the Internet, *American Economic Review* 91, 313–317.
- Lux, T., 2011, Sentiment dynamics and stock returns: the case of the German stock market, *Empirical Economics* 41, 663–679.
- Lv, D., and W. Wu, 2018, Margin trading and price efficiency: information content or price-adjustment speed?, *Accounting and Finance*, <https://doi.org/10.1111/acfi.12403>
- Neal, R., and S. M. Wheatley, 1998, Do measures of investor sentiment predict returns?, *Journal of Financial and Quantitative Analysis* 33, 523–547.
- Qian, L., M. Li, and Y. Li, 2018, Does news travel slowly before a market crash? The role of margin traders, *Accounting and Finance*, <https://doi.org/10.1111/acfi.12419>
- Resnick, P., and R. Zeckhauser, 2002, Trust among strangers in internet transactions: empirical analysis of eBay's reputation system, in: M. R. Baye, ed., *The Economics of the Internet and E-Commerce*, Vol. 11 (Springer, Amsterdam), 127–157.
- Siganos, A., E. Vagenas-Nanos, and P. Verwijmeren, 2014, Facebook's daily sentiment and international stock markets, *Journal of Economic Behavior and Organization* 107, 730–743.
- Solt, M. E., and M. Statman, 1988, How useful is the sentiment index?, *Financial Analysts Journal* 44, 45–55.
- Sprenger, T. O., A. Tumasjan, P. G. Sandner, and I. M. Welp, 2014, Tweets and trades: the information content of stock microblogs, *European Financial Management* 20, 926–957.
- Tetlock, P. C., 2007, Giving content to investor sentiment: the role of media in the stock market, *The Journal of Finance* 62, 1139–1168.
- Tumarkin, R., and R. F. Whitelaw, 2001, News or noise? Internet postings and stock prices, *Financial Analysts Journal* 57, 41–51.
- Wysocki, P., 1998, Cheap talk on the web: The determinants of postings on stock message boards, Working paper (Business School, University of Michigan).
- Xiong, X., Y. Gao, and X. Feng, 2017, Successive short-selling ban lifts and gradual price efficiency: evidence from China, *Accounting and Finance* 57, 1557–1604.
- Xu, H. C., and W. X. Zhou, 2018, A weekly sentiment index and the cross-section of stock returns, *Finance Research Letters* 27, 135–139.
- Yin, Y., and B. Tan, 2017, Analyst's ability, media selection and investor interests: evidence from China, *China Finance Review International* 7, 67–84.
- Zhang, Y., 2009, Determinants of poster reputation on internet stock message boards, *American Journal of Economics and Business Administration* 1, 114–121.
- Zhu, Y., Z. Wu, H. Zhang, and J. Yu, 2017, Media sentiment, institutional investors and probability of stock price crash: evidence from Chinese stock markets, *Accounting and Finance* 57, 1635–1670.

Appendix

Sentiment classification based on Naïve Bayesian Algorithm

Because of the large number of online messages in our sample, we cannot manually classify the sentiment of all the messages. We employ the Natural

Language Process method to classify the messages into three types: positive, neutral and negative.

Chinese sentiment classification consists of two steps: word segmentation and sentiment classification. Word segmentation is usually unnecessary in English sentiment classification because words in English sentences are separated naturally. However, Chinese sentences are composed of Chinese characters. Characters usually have meaning in themselves, e.g. ‘好’ means good. But generally, a word may be composed of 2–4 characters, and the meaning of the characters is not always same as that of the word. For example, the word ‘多头’ means bull side of the market, but neither the character ‘多’ nor ‘头’ has the same meaning. Therefore, we must divide the sentences into the words appropriately and the meaning must remain the same. We employ the FudanNLP-1.6.1 software as our word segmentation instrument, which is also widely used in other studies of natural language processing for Chinese text.

A key factor for word segmentation and sentiment classification is the sentiment dictionary. The dictionary included in FudanNLP-1.6.1 contains a large number of simple and basic words. However, for our study, the dictionary we need must contain more professional words that relate to the finance and stock market context. Therefore, we build a special dictionary for Chinese online stock message boards and classify the messages using the following steps.

Choosing the training sample

We randomly choose 5,000 messages which contain more than five Chinese characters from our online messages sample. The five-character limit is used because short messages lack sentiment content and have no value for training.

Dictionary building and word segmentation of training sample

We segment our 5,000 training sentences using our dictionary, which includes the default dictionary of FudanNLP, HowNet Chinese sentiment dictionary, 219 items of stock market terminology in MBAlib, and all stock names on the Chinese stock market. HowNet is a frequently used dictionary in the study of Chinese sentiment classification (Dong and Dong, 2003) and MBAlib is the biggest encyclopedia website on economics and management in China.

Manual classification

The training sample is classified manually by 10 masters students who are majors in finance and have experience in stock trading. We ask the students to classify the messages into three different types: positive, neutral and negative, and choose the key words which support their judgements. For example, a sentence ‘中行的跳水原因 (the reason for the BOC price collapse)’ is divided

into words as ‘中行 (BOC), 的 (for the), 跳水 (price collapse) and 原因 (reason)’. The word ‘跳水’ (price collapse) is chosen by the student as the key word which makes the sentence sentiment negative.

In order to exclude human error, each message is classified by three different students. When a message is classified as positive and negative by different students, we remove it from our sample. Ten messages were removed and most of our sample was classified into the same type. In particular, when a message is classified as positive/neutral or negative/neutral, we choose the majority one. If the result of all three students is identical, the message will be regarded as the type directly.

Naive Bayesian classification

We collect the sentiment key words identified by the students and remove the meaningless words. Finally, 1,043 words are considered as the Key Sentiment Words of the Chinese stock message board. We employ Naive Bayesian Classification (NBC) for sentiment classification.

The NBC assumes that the occurrences of words are independent of each other. The conditional probability that one message contains the keyword W_I , which is included in Key Sentiment Words and belongs to the sentiment group T_c , $C \in \{Positive, Neutral, Negative\}$, is:

$$P(T_c|W_I) = \frac{P(W_I|T_c)P(T_c)}{P(W_I)} = \frac{P(T_c) \prod_{k=1}^I P(w_k|T_c)}{\prod_{k=1}^I P(w_k)},$$

where w_k is the k th word from the sequence W_I . I is the total number of W_I . Based on Equation 3.1, we can calculate the sentiment probability of each message and we choose the one with the maximum probability as its sentiment type.

$$Type(W_I) = \text{Max}\{P(T_c|W_I)\}, C \in \{Positive, Neutral, Negative\}$$