Personal and Ubiquitous Computing (2021) 25:587–599
https://doi.org/10.1007/s00779-021-01542-3

ORIGINAL ARTICLE

What investors say is what the market says: measuring China's real investor sentiment

Yunchuan Sun[1] · Xiaoping Zeng[1] · Siyu Zhou[2] · Han Zhao[2] · Peter Thomas[1,3] · Haifeng Hu[1]

Received: 16 November 2020 / Accepted: 16 February 2021 / Published online: 26 February 2021

Abstract

This paper describes a novel approach to measure individual investor sentiment using text-based analysis of millions of posts extracted from Chinese financial online forums. We describe how we built a database of more than 200 million stock posts from online financial forums, created *GubaLex*, a sentiment dictionary consisting of 48,878 words to allow sentiment analysis, and how we developed *GubaSenti*, an individual investor sentiment index for the stock market in China. This allowed (1) the first systemic measurement of individual investor sentiment in China; (2) an approach to text-based analysis that reflects investor sentiment about millions of posts about stocks listed in *Guba*; (3) a way to flexibly measure investor sentiment of a single stock, a sector or an industry and the whole market; and (4) made this possible for daily, weekly, monthly, quarterly, and yearly time periods. We also examine the relationship of the sentiment proxy and stock returns and compare it with two typical BW metrics in China. Empirical results show that *GubaSenti* correlates better with market performance than BW metrics in China and can be used to predict market changes in the short term.

Keywords Big data · Individual investor · Investor sentiment · Online financial forums · Sentiment dictionary · Text analysis

1 Introduction

Bubbles, or crashes, are either correlated with extreme optimism (high sentiment) or pessimism (low sentiment), such as the late 1990s bubble in technology stocks in the USA and more recently the coronavirus pandemic that has caused panic among global investors and has led to sharp fluctuations in global stock markets.

In such 'black swan' events such as the pandemic, investor sentiment is more explanatory than other measured. This is supported by work in behavioral finance that asserts that investor sentiment is the main factor that affects investment decisions. Black [1] discusses why 'noise' can cause market inefficiency, Delong et al. [2] describe how sentiment-driven noise trading can lead to mispricing and excess volatility, and Baker and Wurgler [3] present evidence that investor sentiment may have significant effects on the cross-section of stock prices. Other studies have shown that investor sentiment can significantly impact stock returns (see, for example, [4–7]) and have also demonstrated the role of investment sentiment in asset pricing and explaining market anomalies such as the value premium, the momentum effect, and analyst forecast errors (see, for example, [8–11]).

Of course, in order to understand the effect of the impact of investor sentiment on the stock market, it needs to be measured.

As we detail below, there are four main approaches: survey-based, social issue-based, market-based, and text-based.

Survey-based measures are collecting investors' optimistic or pessimistic expectations for the stock market through surveys, such as the University of Michigan Consumer Sentiment Index [12] and the UBS/GALLUP Index for Investor Optimism [13].

✉ Yunchuan Sun
yunch@bnu.edu.cn

Xiaoping Zeng
zengxp@mail.bnu.edu.cn

Peter Thomas
peter.thomas@theorica.io

Haifeng Hu
huhaifeng@bnu.edu.cn

[1] Business School, Beijing Normal University, Beijing 100875, China

[2] School of Mathematical Sciences, Beijing Normal University, Beijing 100875, China

[3] THEORICA, Melbourne, Australia

Social issue-based measures use social events as proxies of investor sentiment, such as aviation disasters [14] and political events [15].

Market-based sentiment metrics measure investor sentiment by market indicators such as trading volume, dividend premium, and initial public offering first-day returns. These are an indirect reflection of investor sentiment and limited by the update of market indicators [16]. Of the market-based measures, the most influential is the widely cited *BW investor sentiment index* [3, 5] that is constructed by principal component analysis of six market-based proxies.

Text-based measures of investor sentiment are made possible through the emergence of online forums for interaction and state-of-the-art tools for textual analysis. They use three main kinds of textual data to measure investor sentiment: *search records from search engines* like *Google* and *Baidu* [16, 17], financial *newspapers* [18], and *data from social media and online financial forums* like *Facebook* in the USA and *Guba* in China [19–21]. Compared with market-based metrics, text-based measures are available in real-time, provide larger data volumes, and, we argue, more directly reflect investor sentiment.

There have been a variety of proposals for text-based sentiment metrics. But as yet, there is no systemic and widely used text-based investor sentiment index. This is because there is no well-developed sentiment dictionary or textual data set available for text-based sentiment analysis, especially in China (but see Sun et al. [22] work to create a lexicon for Chinese *Guba*).

Studies that do exist on the impact of investor sentiment on the stock market are inconclusive. Some studies suggest that investor sentiment is predictive, plays a role in asset pricing, and explains anomalies, whereas others disagree: for example, Devault et al. [23] suggest that market-based metrics like BW index reflect *institutional* investor sentiment in aggregate. It is likely that textual data from online forums reflect the individual investor sentiment and institutional and individual sentiment will have different impacts on the market. And of course, in different countries, the proportions of institutional and individual investors are also different: individual investors in the USA account for about 6% of market value while institutional investors account for more than 93%, meaning that individual investor sentiment will have a limited impact on the market. In China, however, individual investors account for more than 90% of both the volume and frequency or transactions.

Against this background, we built a database of more than 200 million posts from online financial forums like *Guba* in China, and a sentiment dictionary including 48,878 words named *GubaLex* for text-based sentiment analysis. We then developed a measure to assess Chinese individual investor sentiment called *GubaSenti*, the China Individual Investor Sentiment Index.

Gubasenti is notable in four ways: (1) it is the first sentiment index that reflects individual investor sentiment in China; (2) it captures investor sentiment directly based on text-based analysis of data from millions of stock posts in *Guba*; (3) it is flexible and can cover a single stock, stocks of different sectors or industries, the whole market, and any other stock combinations; and (4) it operates in real-time and can focus on different time periods including, but not limited to, hourly, daily, and weekly. Later, in this paper, we will assess the effectiveness of *Gubasenti* compared with two typical BW metrics used in China. Our results indicate that *Gubasenti* correlates better with market performance than BW metrics and can be used to better predict market performance in the short term.

With the emergence of big data technology and online financial forums, investor sentiment can be obtained in real time and more directly. The method and investor sentiment index we propose here is a novel approach that has theoretical and practical relevance.

Firstly, it can be used to test hypotheses emerging from behavioral finance. Most previous works looking at market-based proxies like BW metrics assume that individual investors are responsible for sentiment-induced mispricing, while Devault et al. [23] document that most market-based metrics including BW index capture the demand shocks of institutional, rather than individual investors. *GubaSenti* is a proxy for individual sentiment for the Chinese stock markets and so can be used to re-evaluate these earlier findings. Secondly, the construction of an *individual* investor sentiment index makes it possible to quantify and classify the specific influence of institutional and individual investor sentiment on the stock market, especially in China. This could stimulate further research into real-time sentiment-based asset pricing and quantitative investment. Thirdly, this work can be useful in predicting short-term market performance and making trading strategies, and so help policymakers to better understand the role of individual investors in the stock.

2 Background

2.1 The theoretical basis of investor sentiment

The efficient market hypothesis holds that stock prices equal the rationally discounted value of expected cash flows, and even if there are irrational investors, their demands would be offset by arbitrageurs and ultimately have no significant impact on prices.

However, there are a number of financial anomalies in the market that cannot be explained by the efficient market hypothesis such as underreaction and overreaction, momentum, and the reversal effect. Researchers in behavioral finance have therefore been attempting to develop alternative models built on investor sentiment and investor behavior.

Delong et al. [2] point out that investors are subject to sentiment and discuss the noise trader risk in the financial market and its impact on asset pricing. The systematic deviation of noise traders' expectation of asset value can be regarded as investor sentiment. Lee et al. [24] defines investor sentiment as the part of investors' future return on assets that cannot be explained by fundamentals, that is, the unreasonable deviation in the future price of assets is caused by investor sentiment. Barberis et al. [25] define investor sentiment as the irrational preference of market participants. Baker and Wurgler [3] put forward two definitions of investor sentiment: the propensity to speculate and optimism or pessimism about stocks. Actually, most of extant studies related with investor sentiment are mainly following Delong et al. [2] and Baker and Wurgler [3].

2.2 Measurement of investor sentiment

How to measure investor sentiment is a fundamental challenge before it is possible to quantify its impacts on stock prices. Four main approaches for measurement of investor sentiment are summarized here.

The first is to collect investors' optimistic or pessimistic expectations for the stock market through surveys, such as the University of Michigan Consumer Sentiment Index [12] and the UBS/GALLUP Index for Investor Optimism [13].

The second is to use social events as a proxy of investor sentiment, such as aviation disasters [14] and political events [15].

The third is to measure sentiment based on market indicators, for example trading volume [26], dividend premium [27], and initial public offering first-day returns [28]. So far, the most influential measure is the BW investor sentiment index developed by Baker and Wurger [3, 5]. The BW index is constructed by principal component analysis of six market-based proxies, including trading volume, dividend premium, the closed-end fund discount, the number and first-day returns on IPOs, and the equity share in new issues.

Market-based sentiment metrics like the BW index capture the performance of the market, which may not be investor sentiment but is the *behavioral results* of investor sentiment. A limitation is that they are only available for the overall market, monthly, and lag as they are limited by the frequency with which market indicators are updated. These market-based proxies for investor sentiment are selected to capture the behavior of individual investors following the assumption that individual investors are responsible for sentiment-induced mispricing. However, Devault et al. [23] demonstrate that those sentiment metrics capture the demand shocks of institutional, rather than individual investors.

With the emergence of big data technology and online financial forums, constructing sentiment proxies based on the analysis of text-based data from social network platforms had become a popular way to measure investor sentiment in real time, and more directly than market-based metrics.

There are three types of text-based data available. (1) search records from a search engine. Da et al. [16] constructed a new investor sentiment index using Google search records. Their results show that the index is related to the direction of short-term stock returns and volatility. Fang et al. [17] selected Baidu search records as the proxy of investor sentiment to predict Chinese stock market volatility. (2) *Financial newspapers.* For example, Garćıa [18] used the proportion of positive and negative words in the two-column financial news of the New York Times to measure investor sentiment.(3) *Text-based data from social media and online financial forums*, such as *Twitter* [19, 29, 30], *Facebook* [31], *StockTwits* [32], *Sina Guba*[1] [20, 21], and *Eastmoney Guba*[2] [20, 21]. Compared with survey-based and market-based sentiment proxies, text-based measures are available with much higher frequency and with larger data volume. We would argue that text-based measures are a more direct reflection of investor sentiment, without relying on equilibrium market quantities which may be confounded by a variety of market factors [16, 21].

We feel that there is not yet a novel, effective, and systemic text-based investor sentiment index. One major limitation of using text-based measures as assess Chinese investor sentiment is that the data from online forums like *Guba* in China are colloquial, unlike official announcements, and so there is no effective sentiment dictionary or text-based data set available for Chinese sentiment analysis (but see [22]).

2.3 Effects of investor sentiment on stocks

In respect to the impact of investor sentiment on stock returns, most empirical studies have shown that investor sentiment has a significant impact (see, for example, [4–7, 24]).

Studies have examined whether investor sentiment can predict future returns, but there is no widespread agreement. Wheatley et al. [33] document that the change of investor sentiment would not cause a change of yield but would impact volatility of prices. Kling and Gao [34] demonstrate that Chinese institutional investor sentiment does not have the ability to predict stock market returns either in the short term or in the long term. Xu and Zhou [35] constructed a comprehensive investor sentiment for different portfolios and found that sentiment can predict stock returns in the short term. Nisar and Yeung [36] show that it might be promising to use the sentiment analysis of Twitter data to predict market trends. Sun et al. [37] provide empirical evidence that investor sentiment is more explanatory than other measured as for the

[1] An online financial forum in China, available at http://*Guba*.sina.com.cn/.

[2] An online financial forum in China, available at http://*Guba*.eastmoney.com/.

impact of COVID-19 on the Chinese stock market. Other research has looked at the role of investor sentiment in asset pricing and explaining market anomalies such as the value premium [8] and the momentum effect [9]. These studies are commonly interpreted as reflecting how individual investors' direct trading behaviors impact markets, while Devault et al. [23] demonstrate that the BW sentiment metric captures institutional investors' demand shocks and these earlier findings should be reconsidered.

There are two reasons for the differing results in the research. One is that the *measure* of investor sentiment is different. As shown in Devault et al. [23], market-based measures like the BW metric reflect *institutional* investors' sentiment in aggregate. Relatively speaking, text-based measures based on data from online forums directly reflect the *individual* investor's sentiment. The other reason is that all stock markets are not the same. Taking the USA and China for example, proportions of institutional and individual investors are totally different. In USA, individual investors account for about 6% of the market value only while institutional investors account for more than 93%; hence, individual investor sentiment may have a very limited impact on the market. In China, however, individual investors account for more than 90% in both the number and transaction frequency; thus, they cannot be overlooked and it needs to pay more attention to China's individual investor sentiment.

Hence, we focus on constructing the China Individual Investor Sentiment Index (*Gubasenti*) in this work. It is a direct reflection of individual investor sentiment based on textual analysis of textual data from millions of stock posts in *Guba*, flexible to be derived for single stock, stocks of different sectors or industries, the whole market and any other stock combination, and available at different frequencies, including but not limited to hourly, daily, and weekly.

3 Data and approach

3.1 Introduction to Guba

Guba is an online financial forum for stock information exchange and interaction among investors in China.

In *Guba*, millions of individual investors share news, opinion, and experiences through many, and short, stock posts. Posting, reading, and replying in *Guba* provide an insight into investor sentiment in real time. Unlike formal news and official announcements, these short texts are colloquial and sentiment-enriched.

Guba has also created sub-*Guba*s for each stock which allow investors to interact with each other about a specific stock, so text-based data in stock posts in *Guba* are both market-oriented and individual stock-oriented. The characteristics of *Guba* texts are shown in Table 1.

3.2 Collection of data

We built a distributed web spider to automatically extract stock posts from Sina *Guba* and Eastmoney *Guba* and constructed a large text database.

As of January 2020, there are more than 200 million posts in our database, with 10 trillion replies, as shown in Fig. 1. These posts correspond to more than 1.5 million users with a specific ID and many more anonymous users.

The information about each post includes the stock code, the posting time (accurate to minutes), the type and title of post, the content in the post, the number of post views, any post comments and post sharers, and other user-oriented information, as shown in Table 2.

3.3 Construction of *GubaLex*

Text-based data from posts in *Guba* are sentiment-rich and may reflect individual investors' willingness on a specific stock. While little research has been done related to *Guba*, many other studies look investor sentiment analysis in finance market. Text mining is the most popular method to analyze sentiment which reply on *Lex*icons such as *HowNet*[3] and *NTUSD*[4]. *HowNet* is a comprehensive and systematic commonsense knowledge base and *NTUSD* (developed by National Taiwan University) is an authoritative semantic lexicon. But with the increasing occurrence of new, and highly descriptive, words driven by the rapid development of the Chinese stock market and the growth of the internet, traditional lexicons are difficult to use with data from *Guba*.

So, to allow in-depth analysis of the investor sentiment in *Guba*, we constructed a specified lexicon, named *GubaLex* (see more from the prior work of our team in [22]).

We chose *HowNet* and *NTUSD* as our basic sentiment lexicon and obtained 47,274 sentiment words after eliminating some unfamiliar words manually. We then collected 1434 high-frequency words and stock terms from financial news, stock comments, information about Chinese A-share listed companies, and especially those in our *Guba* corpus, with *JIEBA*[5] which is used for word segmentation.

We then identified 151 *degree words* and 19 *negative words Degree words* that are used to qualify or modify sentiment words in order to make the expression more precise and subtle. Occurrences of *negative words* in stock posts would be counted as double negative and multiple negatives may lead to totally different meanings. An auto update module was built to recognize new words in new posts based on *PAT-Array*, a retrieval algorithm. Table 3 shows the composition of *GubaLex*.

[3] http://www.keenage.com/

[4] http://nlg.csie.ntu.edu.tw/download.php

[5] https://github.com/fxsjy/jieba

Table 1 The characteristics of *Guba* text

Characteristics	Descriptions
Large volume	More than ten thousand posts are posted in *Sina Guba* and *Eastmoney Guba* every day for over 3000 stocks in Chinese listed Companies.
Colloquial	Text in *Guba* are more colloquial than formal news and official announcements from government, companies, analysts, and economists. Investors share news and opinions in a casual and relaxed way.
Individual stock-oriented	*Guba* has created *sub-Gubas* for each stock to allow investors to interact with each other about a specific stock.
Sentiment-enriched	Posts in *Guba* contain more strong sentiments than ordinary financial texts. For example, investors may post “牛!涨涨涨” (*Great! Keep on rising*!) when the stock price continually rises, while “完蛋了!”(*It's over!*) if the shares go down.
Short texts	Investors prefer to express their opinions about some stocks with only several words but with strong sentiments.
Market-oriented	Users of *Guba* are almost all investors trying to obtain or share relevant information by reading or creating posts.
Highly interactive	By posting, reading, and replying in *Guba*, investors can express their feelings and opinions at any time and any place. Interaction the spread of information is faster and more convenient.

3.4 Measurement of sentiment

3.4.1 Investor sentiment of a single post

Each post contains a title, content, and comments. The title, a short string of text, usually represents the point of view of the poster and is a direct reflection of the poster’s sentiment. The post content expands on the title. Comments vary in length and can often contain information such as advertisements, and so, we only analyze the title text.

Each post title may contain positive words, negative words, and a degree adverb. We define the sentiment vector [P, N] of each post as a two-dimensional array. The value of P represents the positive sentimental degree of the post, while the value of N represents the negative sentimental degree of the post. We divide the text into different types of words according to *GubaLex*. Assume there are n sentimental words (bullish words and bearish words) in the post, denoted as w_1, w_2, …, w_n.

At first, the weight of each sentimental word is assigned to 1, and then, it is adjusted according to the following rules: (1) if a k degree adverb ($k = 1, 2, 3, 4$) is used to modify the sentimental word w_i, then the weight of w_i is denoted as k. This is because adverbs have a reinforcing effect on the expression of sentimental words, and adverb level is higher, the reinforcing effect on sentiment is more obvious; (2) if there are an odd number of negative words before the sentimental word w_i, the sentiment will change, and we will regard the positive sentiment as negative sentiment; and (3) If there is an exclamation point at the end of the text, the weight of these sentimental words will double. Finally, the degree of positive and negative sentimental words is summed to obtain the sentimental vector as follows:

$$P = \sum a_i w_i^{(\text{positive})} \tag{1}$$

$$N = \sum b_j w_j^{(\text{negative})} \tag{2}$$

where a_i, b_j represents the adjustment coefficient of sentimental words according to the above rules, $w_i^{(\text{positive})}$, $w_j^{(\text{negative})}$ represents the initial weight of bullish words and bearish words, respectively, with the value of 1. For example, the sentence “I am optimistic about this stock, it must rise today, and will fall tomorrow” contains two bullish words “optimistic” and “rise”, one bearish word “fall” and one second degree

Fig. 1 Stock posts from January 2010 to January 2020

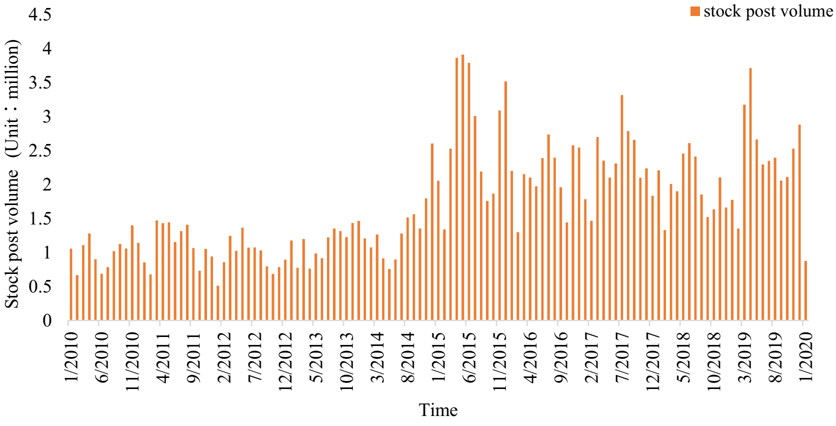

Table 2 The information of each stock post extracted from *Guba*

	Variable name	Description	Data type	Example
Post-oriented	Code	Stock code	Integer	600519
	Time	Posting time	Time	2020/1/7 16:43
	Type	Type of post	String	“热帖”(Hot post) /“普通帖”(Normal post)
	Title	Title of post	String	“等着分红” (Waiting for dividends)
	Content	Content in the post	String	“继续关注小市值老牌绩优股...” (Keep focusing on stocks of less market value but good performance...)
	View	Number of post views	Integer	1936
	Comment	Number of post comments	Integer	13
	Share	Number of post sharing	Integer	5
User-oriented	Name	Name in *Guba*	String	Shuchao168
	Age	Net age in *Guba*	Float	6.3 years
	Influence	Influence index in *Guba*	Float	4.5 (range from 0.0–5.0)

adverb “must” which modifies “rise”, so the final sentiment vector is $[P, N] = [1 + 2 \times 1, 1] = [3, 1]$.

Then, we calculate the impact factor of the post. The number of reading and replied comments implies the post’s influence on other investors. Posts with more readings and comments tend to have larger influence, so the influence factor of such posts is greater. Generally, the number of reading of posts is often much higher than that of comments. If a stock has a total of n_t posts during the period of t, and each post i also includes the number of reading $r_{i,\ t}$ and the number of comments $c_{i,\ t}$, then we define the influence factor as follows:

$$h_{i,t} = \frac{\sum_{i=1}^{n_t} r_{i,t}}{\sum_{i=1}^{n_t} c_{i,t}} \times c_{i,t} + r_{i,t} \tag{3}$$

The influence factor is further smoothed to the vicinity of 1, as follows:

$$H_{i,t} = \ln\left(h_{i,t}\right) - \frac{\sum_{i=1}^{n_t} \ln\left(h_{i,t}\right)}{n_t} + 1 \tag{4}$$

The third step is to calculate the sentiment index of single post. Sentiment index should include not only the sentiment vector information contained in the title text of the post, but also the influence factor of the post. Therefore, the sentiment index of post i is

$$T_{i,t} = \chi \times \ln\left(1 + |P_{i,t} - N_{i,t}|\right) \times H_{i,t} \tag{5}$$

where χ represents whether the post is positive or negative. If the sentimental vector $[P_{i,\ t}, N_{i,\ t}]$ of the post satisfies $P_{i,\ t} > N_{i,\ t}$, which means the post sentiment is positive, then χ is 1, otherwise χ is − 1. This sentiment index evaluation method not only takes into account the sentimental expression and intensity of the sentiment, but also considers the external influence such as the number of reading and comments of the post. It is a relatively comprehensive and reasonable evaluation method.

3.4.2 Investor sentiment of a single stock

In the prior work of our team (see [20]), we use the method introduced by Antweiler and Frank [29] to synthesize sentiment index of a single stock as follows:

$$S = ln\frac{1 + N_p}{1 + N_n} \tag{6}$$

where N_p and N_n represent the number of positive posts and negative posts, respectively.

However, this method has several disadvantages. Firstly, it cannot reflect the sentiment differences among different posts. This measure only considers the number of positive and negative posts but ignores the text of the posts. In fact, the more

Table 3 The composition of *GubaLex*

Type of word		Number	Example
Basic words	Bullish words	5782	“好” (Good)
	Bearish words	10725	“差” (Bad)
	Others	30767	“正常” (Normal)
Stock terms	Bullish words	821	“给力” (Awesome)
	Bearish words	613	“破位” (Break down)
Degree words	Level 1	30	“略微” (Slightly)
	Level 2	40	“相对地” (Relatively)
	Level 3	37	“非常” (Very)
	Level 4	44	“极其, 完完全全” (Extremely)
Negative words	Negative	19	“不曾” (Never)
Total		48878	

bullish words the post has, the stronger the positive sentiment of investors are, and vice versa. Secondly, it cannot reflect the impact of the number of posts with regard to the targeted stock. When the number of posts N_{p} and N_{n} is much larger than 1, the sentiment index of the targeted stock is only related to the relative ratio of positive and negative posts. Here, we will improve this method as detailed below.

The number of active users and posts are different among *sub-Gubas* for each stock. Usually, popular stocks have more active users and posts. In order to fully reflect the impact of the number of posts on investor sentiment of each stock, we construct the sentiment index of a single stock by integrating the sentiment index of each post as follows: suppose a stock has n_{t} posts in period t , and the sentiment index of post i is $T_{\mathrm{i,\ t}}$, then the sentiment index of the stock in period t is

$$E_{\mathrm{t}} = \frac{ln(n_{\mathrm{t}})}{n_{\mathrm{t}}} \sum_{i=1}^{n_{\mathrm{t}}} T_{\mathrm{i,t}} \tag{7}$$

3.4.3 Investor sentiment for a set of stocks

Stocks in China can be divided into Shanghai, Shenzhen, and Hong Kong stocks by stock exchange and divided into finance stocks, real estate stocks, agriculture stocks, and so on by industry.

We define a *stock set* as a collection of stocks classified in a certain way, and the sentiment index of the stock set is related to the sentiment index of stocks belonging to the set. We construct the sentiment index of a stock set as follows: suppose the stock set include k stocks $s_1, s_2, \ldots, s_{\mathrm{k}}$. The sentiment index of stock s_{j} in the period of t is $E_{\mathrm{j,\ t}}$, and the stock has $n_{\mathrm{j,\ t}}$ posts in the period of t, then the sentiment index of the stock set in the period of t is

$$HE_{\mathrm{t}} = \frac{ln\left(\sum_{j=1}^{k} n_{\mathrm{j,t}}\right)}{\sum_{j=1}^{k} n_{\mathrm{j,t}}} \sum_{j=1}^{k} E_{\mathrm{j,t}} \tag{8}$$

4 Verification

Using the method above, individual investor sentiment could be derived for a single stock, for industries and the overall market at different frequencies including but not limited to hourly, daily, weekly, and monthly.

For simplicity, we refer to the individual investor sentiment index based on *Guba* text-based analysis in the following as *GubaSenti*. To verify the effectiveness of *Gubasenti*, we will first examine the correlation between *Gubasenti* and CSI 300 Index (*CSI300*)[6], then compare *Gubasenti* with two popular market-based investor sentiment index in China (the BW metrics in China), and several typical text-based measures in extant literature.

4.1 GubaSenti and stock returns

4.1.1 Market-oriented GubaSenti and CSI 300 index

We analyzed the textual data of a total of 27,503,441 posts in *Guba* of 300 stocks involved in *CSI300* from February 8, 2013 to December 27, 2019. There are 15,381,281 posts with sentiment identified by *GubaLex* accounting for 56.0% of the whole posts, and the remaining posts are recognized as without sentiment text or with invalid information such as advertisements.

Figure 2 illustrates the market tendency of *Gubasenti* and *CSI300* from February 8, 2013 to December 27, 2019. The weekly *Gubasenti* of 300 stocks involved in *CSI300* is mostly distributed in the range of 0 to 20,000, and only a small part of the sentiment is negative.

It can be seen from Fig. 2 that *Gubasenti* has an obvious positive correlation with *CSI300* overall. For example, in the week from April 20, 2015 to April 24, 2015, *Gubasenti* reached a peak value of 73,652 points, and *CSI300* rose by 2.3% in the same week with a consecutive rise for 7 weeks, while in the week from July 27, *Gubasenti* reached a negative peak value of − 18,234 points, and *CSI300* fell by 8.6% during the week.

Gubasenti and *CSI300* are in step in most cases provides a preliminary indication that *Gubasenti* reflects the performance of the stock market.

We also did further analysis on the correlation between *Gubasenti* and *CSI300*. Table 4 shows that there is a significant positive correlation between *Gubasenti* and *CSI300*. Specifically, the Pearson correlation coefficients between *Gubasenti* of a certain week and *CSI300* of the same week, the last week and the next week are 0.56, 0.34, and 0.15, respectively, all significant at 1% level.

Moreover, we did a regression analysis on the relationship between *Gubasenti* and *CSI300* with the following model:

$$CSI300_{\mathrm{t+k}} = a_0 + a_1 GubaSenti_{\mathrm{t}} + a_2 SMB_{\mathrm{t}} + a_3 HML_{\mathrm{t}} + \varepsilon_{\mathrm{t}} \tag{9}$$

SMB and *HML* are factors from Fama-French model, selected here as control variables[7]. *SMB* refers to the market value factor and *HML* refers to book to market ratio factor. The reason for not including *RMKT*, market risk premium factor in Fama-French model, is that there is strong collinearity between *RMKT* and *CSI300*.

[6] CSI 300 Index consists of the 300 largest and most liquid A-share stocks. The Index aims to reflect the overall performance of China A-share market. Available at http://www.csindex.com.cn/en/indices/index-detail/000300.

[7] Data of Fama-French factors are downloaded in CSMER database (China Stock Market and Accounting Research Database). Available at http://www.gtarsc.com/.

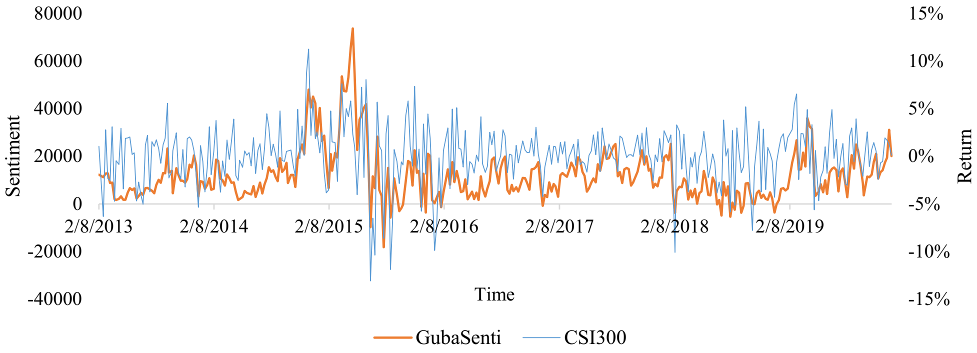

Fig. 2 The tendency of *Gubasenti* and *CSI300* from February 8, 2013 to December 27, 2019

Panel A of Table 5 shows the results of regression between market-oriented *Gubasenti* and *CSI300* without any control variables, and in panel B of Table 5, the results with control variables. It can be seen that the estimated coefficients of *Gubasenti* are positive and significant at 1% level in regressions between *Gubasenti* in week t and *CSI300* in week t and week t + 1. The result is similar with control market variables. The results demonstrate that *Gubasenti* can predict market performance in a short run, and the higher the sentiment, the better the market performance would be.

4.1.2 Single stock-oriented GubaSenti and returns on each stock involved in CSI300

We also examined the correlation between *Gubasenti* and returns of each stock involved in *CSI300.*

There are 280 stocks with comprehensive data from February 2013 to December 2019. The results show that *Gubasenti* of 279 stocks are positively correlated with their returns, significant at 1% level, which means that *Gubasenti* would be effective when applied to analysis of a single stock. Table 6 presents the distribution of Pearson correlation coefficients between *Gubasenti* of 279 stocks and their returns in detail, over 70% exceeding 0.4.

It is found that the Pearson correlation coefficient between *Gubasenti* and stock return is impacted by the post volume of each stock in Guba. Specifically, from February 2013 to December 2019, the mean post of 26 stocks with Pearson correlation coefficients less than 0.3 is 49,460, the mean post of 29 stocks with Pearson correlation coefficients over 0.6 is 135,348, and that of total 280 stocks is 99,415. This result implies that big volume of stock posts is the prerequisite of *Gubasenti* being effective.

Table 4 Pearson correlation coefficients between *Gubasenti* and *CSI300* (t = week)

	$CSI300_{t-1}$	$CSI300_{t}$	$CSI300_{t+1}$
$Gubasenti_{t}$	0.342^{***}	0.557^{***}	0.152^{***}

We further did a regression analysis on the relationship between single stock-oriented *Gubasenti* and stock return of each stock in *CSI300* with the following model:

$$Return_{i,t+k} = a_0 + a_1 GubaSenti_{i,t} + a_2 RMKT_t + a_3 SMB_t + a_4 HML_t + \varepsilon_t, \qquad (10)$$

Fama-French factors are selected to be control variables, where RMKT is the market risk premium factor, SMB is the market value factor, and HML is the book to market ratio factor. Panel A of Table 7 shows the results of regression between single stock-oriented *Gubasenti* and stock return of each stock in *CSI300* without any control variables, and Panel B of Table 7 shows the results with control variables.

It can be seen that the estimated coefficients of *Gubasenti* are positive and significant at 1% level in regressions between *Gubasenti* in week t and return in week *t*, *t* + 1 and *t* + 2. The result is similar when control market variables are added. The results further indicate that *Gubasenti* can predict stock performance in the short term, and *Gubasenti* would be more effective in predicting the performance of single stock than performance of the overall market.

4.2 Comparison between *GubaSenti* and *BW* metrics in China

4.2.1 Comparison in basic features

There are two main market-based sentiment metrics in China, *CICSI* (China investor composite sentiment index) and *ISI* (Investor Sentiment Index)[8], both of which are typical applications of the method proposed by Baker and Wurgler (2006) to the Chinese stock market.

CICSI is based on principal component analysis of six market indicators, including the closed-fund discount, the number on IPOs, the average first-day returns on IPOs, new investor accounts, consumer confidence, and trading volume.

[8] *CICSI* and *ISI* can be downloaded in CSMER database (China Stock Market and Accounting Research Database). Available at http://www.gtarsc.com/#/datacenter/singletable/search?serId=16anddatabaseId=157.

Table 5 Results of regression between *Gubasenti* and *CSI300* (*t* = week)

Panel A. Results of regression between market-oriented Gubasenti and CSI300 (no control variable)					
	$CSI300_t$	$CSI300_{t+1}$	$CSI300_{t+2}$	$CSI300_{t+3}$	$CSI300_{t+4}$
$Gubasenti_t$	0.153^{***}	0.042^{***}	0.022	0.008	0.019
$Adj\text{-}R^2$ (%)	30.99	2.32	0.65	0.11	0.49
Panel B. Results of regression between market-oriented Gubasenti and CSI300 (with control variables)					
	$CSI300_t$	$CSI300_{t+1}$	$CSI300_{t+2}$	$CSI300_{t+3}$	$CSI300_{t+4}$
$Gubasenti_t$	0.152^{***}	0.041^{***}	0.022	0.008	0.019
SMB_t	− 0.035	− 0.119	-0.172^{*}	− 0.066	− 0.053
HML_t	− 0.190	− 0.188	− 0.181	− 0.224	− 0.021
$Adj\text{-}R^2$ (%)	31.73	2.74	1.51	0.88	0.64

Similar to *CICSI*, *ISI* is also based on six market indicators, and the only difference is that trading volume is replaced by turnover. *CICSI* and *ISI* are both relying on market indicators to indirectly measure the investor sentiment, which may be confounded by a variety of market factors. Because of the update frequency of market indicators, *CICSI* and *ISI* can only be obtained monthly. Moreover, *CICSI* and *ISI* are *market-oriented* and so cannot distinguish between *institutional* investor sentiment and *individual* investor sentiment. In addition, both *CICSI* and *ISI* are only available for the investor sentiment of the whole market, not available for a single stock or industries.

Gubasenti outperforms *CICSI* and *ISI* with regard to issues above. Firstly, *Gubasenti* is based on text-based analysis of textual data from stock posts in online financial forums like *Guba*, directly reflecting investor sentiment. Secondly, *Gubasenti* are available with much higher frequency. *Gubasenti* is constructed to reflect individual investor sentiment in China specially, set against the background that individual investors play a vital role in Chinese stock market. Thirdly, *Gubasenti* is flexible to measure investor sentiment of a single stock, a sector or an industry, and the whole market. Table 8 shows the comparison between *Gubasenti* and the two main market-based sentiment metrics in basic features.

The Pearson correlation coefficient between *Gubasenti* and *CICSI* is 0.27, significant at 5% level, and that between *Gubasenti* and *ISI* is 0.64, significant at 1% level, which further indicates that *Gubasenti* would be a complement to market-based sentiment metrics.

Table 6 The distribution of Pearson correlation coefficients between *Gubasenti* of 279 stocks and their returns

Range of coefficients	0–0.1	0.1–0.2	0.2–0.3	0.3–0.4	0.4–0.5
Number	1	3	22	50	108
Percentage	0.36%	1.07%	7.86%	17.86%	38.57%
Range of coefficients	0.5–0.6	0.6–0.7	0.7–0.8	0.8–0.9	0.9–1.0
Number	67	26	3	0	0
Percentage	23.93%	9.29%	1.07%	0.00%	0.00%

4.2.2 Comparison in the correlation with stock returns

Figure 3 illustrates the tendency of *CICSI*, *ISI*, *Gubasenti*, and *CSI300* at monthly frequency from February 2013 to December 2019[9].

It can be seen that *Gubasenti* is more consistent with *CSI300* both in tendency and fluctuations, compared with *CICSI* and *ISI*. Further, we can examine the Pearson correlation coefficients between *CICSI*, *ISI*, *Gubasenti*, and *CSI300* respectively, as shown in Panel A of Table 9.

The Pearson correlation coefficients between *CICSI* of a specific month and *CSI300* of the same month, the last month and the next month are 0.12 (not significant), 0.26 (significant at 5% level), and 0.016 (not significant), respectively; the Pearson correlation coefficients between *ISI* of a specific month and *CSI300* of the same month, the last month and the next month are 0.28 (significant at 5% level), 0.36 (significant at 1% level), and 0.08 (not significant), respectively; and the Pearson correlation coefficients between *Gubasenti* of a certain month and *CSI300* of the same month, the last month and the next month are 0.61 (significant at 1% level), 0.50 (significant at 1% level), and 0.07 (not significant), respectively.

The difference in correlation coefficients with *CSI300* is significant between *Gubasenti* and *CICSI*, and between *Gubasenti* and *ISI*, as shown in Panel B of Table 9.

We can therefore suggest that *Gubasenti* is more effective than *CICSI* and *ISI*. Another meaningful finding is that *Gubasenti* in a specific week is positively correlated with *CSI300* in the *next week*, while the correlation is not significant at monthly frequency, which highlights the effectiveness of *Gubasenti* in predicting market performance in the short term.

[9] *Gubasenti* is preprocessed to be of the same order of magnitude and frequency as *CICSI* and *ISI*.

Table 7 Results of regression between *Gubasenti* and stock returns (t = week)

Panel A. Results of regression between single stock-oriented Gubasenti and return of each stock involved in CSI300 (no control variable)					
	$Return_t$	$Return_{t+1}$	$Return_{t+2}$	$Return_{t+3}$	$Return_{t+4}$
$Gubasenti_t$	0.438^{***}	0.049^{***}	0.016^{***}	0.004	0.004
$Adj\text{-}R^2$ (%)	12.39	0.16	0.02	8.56E-04	8.81E-04
Panel B. Results of regression between single stock-oriented *Gubasenti* and return of each stock involved in *CSI300* (with control variables)					
	$Return_t$	$Return_{t+1}$	$Return_{t+2}$	$Return_{t+3}$	$Return_{t+4}$
$Gubasenti_t$	0.371^{***}	0.047^{***}	0.007^{*}	0.014^{***}	0.002
$RMKT_t$	0.845^{***}	0.104^{***}	0.067^{***}	-0.084^{***}	0.060^{***}
SMB_t	0.040^{***}	-0.198^{***}	-0.250^{***}	0.029^{**}	0.146^{***}
HML_t	-0.244^{***}	-0.279^{***}	-0.258^{***}	-0.173^{***}	-0.175^{***}
$Adj\text{-}R^2$ (%)	31.30	0.71	0.48	0.26	0.20

Table 8 The comparison between *Gubasenti* and two main market-based sentiment metrics in basic features

	Gubasenti	*CICSI*	*ISI*
Data	Textual data from stock posts in online financial forums like *Guba*	The closed-fund discount The number on IPOs The average first-day returns on IPOs New investor accounts Consumer confidence Trading volume	The closed-fund discount The number on IPOs The average first-day returns on IPOs New investor accounts Consumer confidence Turnover
Approach	Text analysis	Principal component analysis	Principal component analysis
Frequency	Hourly, daily, weekly, monthly, etc	Monthly	
Investor targeted	Individual investors	Investors in the whole market	
Object targeted	Single stock, sectors, industries, and the whole market	The whole market	

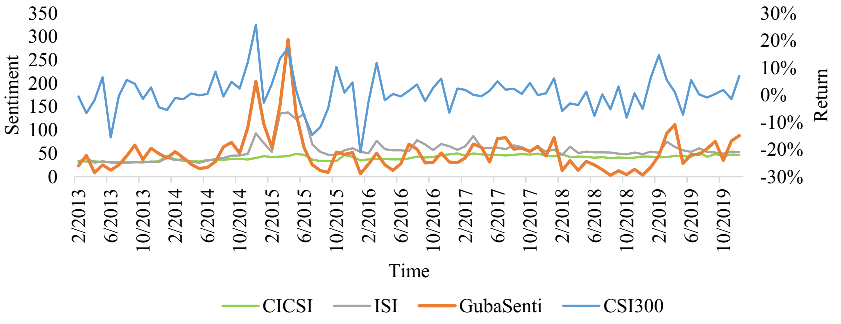

Fig. 3 The tendency of *CICSI*, *ISI*, *Gubasenti*, and *CSI300* from February 2013 to December 2019

4.3 Comparison between *Gubasenti* and other typical text-based metrics

With the emergence and rapid development of the Internet and big data technology, constructing sentiment proxies based on the analysis of text-based data from social network platforms had become more and more popular. As mentioned earlier, there are mainly three kinds of text-based data used in extant researches, including *search records from a search engine, financial newspapers, and text-based data from social media or online financial forums.* Here, we select four typical text-based metrics from extant literature to compare with Gubasenti constructed in our work, as shown in Table 10.

Da et al. [16] constructed a new investor sentiment index using Google search records and showed that the index is related to the direction of short-term stock returns and volatility. Garćıa [18] used the proportion of positive and negative words in the two-column financial news of the New York Times to measure investor sentiment and demonstrated that the predictability of stock returns using news' content is concentrated in recessions. Siganos et al. [31] measured the distance between people with positive and negative sentiment on a daily basis for 20 countries by using data from status updates on Facebook and found that divergence of sentiment is positively related to trading volume and stock price volatility. Renault [32] implemented a novel approach to derive investor sentiment from messages posted on StockTwits and, provided empirical evidence that online investor sentiment helps forecast intraday stock index returns.

It is obvious that four text-based metrics are available with high frequency, at least daily. All of them are designed to measure investor sentiment of the market, without highlighting the individual investors or institutional investors. Actually, in the US market, individual investors account for about only 6% of the market value while institutional investors account for more than 93%; hence, individual investor sentiment may have a very limited impact on the market. Though similar with four metrics in text-based data acquisition and main approach, *Gubasenti* differs from them in three aspects as follows.

First, *Gubasenti* is constructed especially to reflect individual investor sentiment in China as individual investors play a vital role in the Chinese stock market, accounting for more than 90% in both the number and transaction frequency. Second, *Gubasenti* is a systemic investor sentiment index, flexible to measure investor sentiment of a single stock, a sector or an industry and the whole market.

Table 9 Pearson correlation coefficients between three sentiment index and *CSI300* (*T* = month)

Panel A. Pearson correlation coefficients between three sentiment index and *CSI300*			
	$CSI300_{T-1}$	$CSI300_{T}$	$CSI300_{T+1}$
$Gubasenti_{T}$	0.503^{***}	0.609^{***}	0.073
$CICSI_{T}$	0.261^{**}	0.121	− 0.016
ISI_{T}	0.365^{***}	0.275^{**}	− 0.079
Panel B. Difference in Pearson correlation coefficients between three sentiment index and *CSI300*			
	$CSI300_{T-1}$	$CSI300_{T}$	$CSI300_{T+1}$
$Gubasenti_{T}$ *vs.* $CICSI_{T}$	2.095^{**}	4.552^{***}	0.663
$Gubasenti_{T}$ *vs.* ISI_{T}	1.714^{*}	4.569^{***}	1.655^{*}
$CICSI_{T}$ *vs.* ISI_{T}	1..040	1.501	− 0.593

Table 10 The comparison between *Gubasenti* and four typical text-based metrics

	Gubasenti	Other typical text-based metrics			
		Da et al. (2014)	Garćia (2013)	Siganos et al. (2017)	Renault (2017)
Data	Textual data from stock posts in online financial forums like Guba	Search records from Google, a search engine	Financial news from the New York Times, a financial newspaper	Data from status updates on Facebook, a social media	Text-based data from StockTwits, an online financial forum
Stock market	The Chinese stock market	The US stock market	The US stock market	Including 20 countries, like America	The US stock market
Main approach	Text analysis	Text analysis	Text analysis	Text analysis	Text analysis
Frequency	Hourly, daily, weekly, monthly, etc	Daily	Daily	Daily	Hourly
Investor targeted	Individual investors	Investors in the market			
Object targeted	Single stock, sectors, industries, and the whole market	The whole market			

5 Conclusion and discussion

5.1 Conclusion

This paper has described a novel measurement of China's individual investor sentiment using text-based data from online financial forums.

We built a text-based database consisting of more than 200 million stock posts from *Guba* in China and a specific sentiment dictionary including 48,878 words (*GubaLex*) for textual sentiment analysis.

We then proposed a novel method to measure China's individual investor sentiment and constructed *GubaSenti* the China Individual Investor Sentiment Index. We then examined the correlation between *Gubasenti* and stock returns and compared *Gubasenti* with two typical *BW* metrics in China.

The results indicate that *Gubasenti* is more relevant to market performance than market-based sentiment metrics and can be used to predict market performance in the short term.

Compared with traditional sentiment metrics, *Gubasenti*:

(1) is a systematic sentiment index constructed especially to reflect individual investor sentiment in China, set in the context that individual investors play an vital role in the Chinese stock market, accounting for more than 90% in both the number and transaction frequency

(2) captures investor sentiment more directly, using text-based analysis of data from stock posts in online financial forums like Guba, whereas market-based sentiment metrics like *BW* index capture the performance of the market which is not the same as investor sentiment.

(3) is flexible to examine a single stock, stocks of different sectors or industries, the overall market and any other combination. Traditional market-based sentiment metrics like BW index are only available for the overall market.

(4) is available at a much higher frequency than market-based sentiment metrics. *Gubasenti* is real-time at hourly, daily, weekly, and monthly frequencies, while traditional market-based sentiment metrics lag limited by the update frequency of market indicators and could only be obtained monthly.

In conclusion, *Gubasenti* is the first systemic, flexible, real-time representative proxy of China individual investor sentiment using text-based analysis of big data from millions of stock posts in online financial forums.

5.2 Discussion

Gubasenti can contribute both to theory and practice. It can be used to test hypotheses from behavioral finance. *Gubasenti* is a proxy for individual sentiment for Chinese stock markets and can be used to reexamine earlier findings. An individual investor sentiment index makes it possible to quantify the influences of institutional and individual investor sentiment on the stock market, especially for China, which could stimulate research on real-time sentiment-based asset pricing. *Gubasenti* could be useful to product short-term market performance and to inform trading strategies and can help policy makers to better understand the role of individual investors in the stock market.

Real-time sentiment-based asset pricing is a promising area for future research and can play a vital role in quantitative investment in practice.

We would like to share *Gubasenti* dataset for the academic research purpose on the official website of the International Institute of Big Data in Finance based at the Business School, Beijing Normal University. It is available at:

https://ifind.bnu.edu.cn/sjzl/data/, or please contact via email: ifind@bnu.edu.cn.

References

1. Black F (1986) Noise. *J Financ 41*:529–543
2. Delong JB, Shleifer A, Summers LH, Waldmann RJ (1990) Noise trader risk in financial markets. *J Polit Econ 98*:703–738
3. Baker M, Wurgler J (2006) Investor sentiment and the cross-section of stock returns. *J Financ 61*:1645–1680
4. Schmeling M (2007) Institutional and individual sentiment: smart money and noise trader risk? *Int J Forecast 23*:127–145
5. Wurgler J, Baker M (2007) Investor sentiment in the stock market. *J Econ Perspect 21*:129–151
6. Ben-Rephael A, Kandel S, Wohl A (2012) Measuring investor sentiment with mutual fund flows. *J Financ Econ 104*:363–382
7. Ruan Q, Wang Z, Zhou Y, Lv D (2019) A new investor sentiment indicator (ISI) based on artificial intelligence: a powerful return predictor in China. *Econ Model*
8. Piotroski JD, So EC (2012) Identifying expectation errors in value/glamour strategies: a fundamental analysis approach. *Rev Financ Stud 25*:2841–2875
9. Antoniou C, Doukas JA, Subrahmanyam A (2013) Cognitive dissonance, sentiment, and momentum. *J Financ Quant Anal 48*:245–275
10. Stambaugh RF, Yu J, Yuan Y (2015) Arbitrage asymmetry and the idiosyncratic volatility puzzle. *J Financ 70*:1903–1948
11. Liu J, Stambaugh RF, Yuan Y (2019) Size and value in China. *J Financ Econ 134*:48–69
12. Brown GW, Cliff MT (2005) Investor sentiment and asset valuation. *J Bus 78*:405–440
13. Lemmon M, Portniaguina E (2006) Consumer confidence and asset prices: Some empirical evidence. *Rev Financ Stud 19*:1499–1529
14. Kaplanski G, Levy H (2010) Sentiment and stock prices: the case of aviation disasters. *J Financ Econ 95*:174–201
15. Hwang B (2011) Country-specific sentiment and security prices. *J Financ Econ 100*:382–401
16. Da Z, Engelberg J, Gao P (2014) The sum of all FEARS investor sentiment and asset prices. *Rev Financ Stud 28*:1–32
17. Fang J, Gozgor G, Lau CM, Lu Z (2020) The impact of Baidu Index sentiment on the volatility of China's stock markets. *Financ Res Lett 32*:101099
18. Garcia D (2013) Sentiment during recessions. *J Financ 68*:1267–1300
19. Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. *J Comput Sci-Neth 2*:1–8
20. Sun Y, Fang M, Wang X (2018) A novel stock recommendation system using Guba sentiment analysis. *Pers Ubiquit Comput 22*:575–587
21. Li, J.; Chen, Y.; Shen, Y.; Wang, J.; Huang, Z. Measuring China's stock market sentiment. *SSRN, http://ssrn.com/abstract=3377684*. 2019.
22. Sun, Y.; Fang, M.; Wang, X.; Diao, S. GubaLex: Guba-Oriented Sentiment Lexicon for Big Texts in Finance. 13th International Conference on Semantics, Knowledge and Grids (SKG); 2017 2017-08-13; Beijing: IEEE; 2017. p. 25-32.
23. Devault L, Sias R, Starks L (2019) Sentiment metrics and investor demand. *J Financ 74*:985–1024
24. Lee C, Shleifer A, Thaler RH (1991) Investor sentiment and the closed-end fund puzzle. *J Financ 46*:75–109
25. Barberis N, Huang M, Santos T (2001) Prospect theory and asset prices. *Q J Econ 116*:1–53
26. Baker M, Stein JC (2004) Market liquidity as a sentiment indicator. *J Financ Mark 7*:271–299
27. Baker M, Wurgler J (2004) A catering theory of dividends. *J Financ 59*:1125–1165
28. Ljungqvist A, Wilhelm W (2003) IPO Pricing in the Dot-cor Bubble. *J Financ 58*:723–752
29. Antweiler W, Frank MZ (2004) Is all that talk just noise? The information content of Internet stock message boards. *J Financ 59*:1259–1294
30. Shen D, Liu L, Zhang Y (2018) Quantifying the cross-sectional relationship between online sentiment and the skewness of stock returns. *Physica A 490*:928–934
31. Siganos A, Vagenas-Nanos E, Verwijmeren P (2017) Divergence of sentiment and stock market trading. *J Bank Financ 78*:130–141
32. Renault T (2017) Intraday online investor sentiment and return patterns in the U.S. stock market. *J Bank Financ 84*:25–40
33. Neal R, Wheatley SM (1998) Do measures of investor sentiment predict returns? *J Financ Quant Anal 33*:523–547
34. Kling G, Gao L (2008) Chinese institutional investors' sentiment. *J Int Financ Mark Inst Money 18*:374–387
35. Xu H, Zhou W (2018) A weekly sentiment index and the cross-section of stock returns. *Financ Res Lett 27*:135–139
36. Nisar TM, Yeung M (2018) Twitter as a tool for forecasting stock market movements: a short-window event study. *J Finan Data Sci 4*:101–119
37. Sun Y, Wu M, Zeng X, Peng Z (2021) The impact of COVID-19 on the Chinese stock market: Sentimental or substantial?. *Financ Res Lett 38*:101838