MEASURING CHINA’S STOCK MARKET SENTIMENT*

Jia Li†
Yun Chen‡
Yan Shen§
Jingyi Wang¶
Zhuo Huang∥

Abstract
This paper develops textual sentiment measures for China’s stock market by extracting the textual tone of 60 million messages posted on a major online investor forum in China from 2008 to 2018. We conduct sentiment extraction by using both conventional dictionary methods based on customized word lists and supervised machine-learning methods (support vector machine and convolutional neural network). The market-level textual sentiment index is constructed as the average of message-level sentiment scores, and the textual disagreement index is constructed as their dispersion. These textual measures allow us to test a range of predictions of classical behavioral asset-pricing models within a unified empirical setting. We find that textual sentiment can significantly predict market return, exhibiting a salient underreaction-overreaction pattern on a time scale of several months. This effect is more pronounced for small and growth stocks, and is stronger under higher investor attention and during more volatile periods. We also find that textual sentiment exerts a significant and asymmetric impact on future volatility. Finally, we show that trading volume will be higher when textual sentiment is unusually high or low and when there are more differences of opinion, as measured by our textual disagreement. Based on a massive textual dataset, our analysis provides support for the noise-trading theory and the limits-to-arbitrage argument, as well as predictions from limited-attention and disagreement models.

Keywords: disagreement, machine learning, noise trading, sentiment, textual analysis, volatility, volume.

JEL classification: C45, C53, C55, G12, G41.

*We thank Danxu Cheng, Yajing Dai, Hongcheng Li, Xingyu Li, Yabo Li, Fei Qin, Yiming Shen, Zhimin Qiu, Guangxiang Song, Rongyu You, Hanchi Wang, Jieru Wang, Hanyu Zhang and Hong Yan for research assistance, Xiaomeng Du and Wei Huang (Baifendian) for data collection assistance, and Siyu You (Google) for advice on the convolutional neural network method. We also thank Tim Bollerslev, Peter Hansen, Cheng Hsiao, Justin Yifu Lin, Andrew Patton, Minggao Shen, Guoqing Song and Brian Weller for helpful comments. Financial support from the Baifendian Information Technology Co., Ltd. is gratefully acknowledged. This version: April 20, 2019.
†Corresponding author. Department of Economics, Duke University, Durham, NC 27708; e-mail: jli410@duke.edu.
‡National School of Development, Peking University, Beijing, China 100871; e-mail: yunchen@pku.edu.cn.
§National School of Development, Peking University, Beijing, China 100871; e-mail: yshen@nsd.pku.edu.cn.
¶National School of Development, Peking University, Beijing, China 100871; e-mail: wangjy1992@pku.edu.cn.
∥National School of Development, Peking University, Beijing, China 100871; e-mail: zhuohuang@nsd.pku.edu.cn.
I. Introduction

In his 1936 masterwork, *The General Theory of Employment, Interest and Money*, John Maynard Keynes argued that much economic activity is governed by “animal spirits.”\(^1\) Based on experimental evidence, Tversky and Kahneman (1974) articulate a list of cognitive biases in judgmental heuristics. Black (1986) suggests that Kahneman and Tversky’s theory may help describe the motivation of noise traders depicted by Kyle (1985), and discusses why “noise” can cause market inefficiency. De Long et al. (1990) formally demonstrate that sentiment-driven noise trading can lead to mispricing and excess volatility (Shiller (1981)) when rational arbitrageurs face limits of arbitrage (Shleifer and Vishny (1997)). Further, Hong and Stein (2007) emphasize the importance of jointly modeling asset price and trading volume and advocate “disagreement” models in which investors hold different opinions and agree to disagree (Aumann (1976)).

Meanwhile, measuring investor sentiment and disagreement, and quantifying their effects on market activities, are at the center of the related empirical literature (Baker and Wurgler (2007)). In this paper, we conduct measurements using a unique dataset that consists of 60 million text messages posted on a major online investor forum in China. Relying on state-of-the-art tools from computational linguistics, we extract the textual tones of these messages and use their average and dispersion within each period to measure the corresponding market-level sentiment and disagreement, respectively. These textual measures allow us to test a range of hypotheses from the aforementioned theoretical literature within a unified empirical framework for China’s stock market. In his presidential address at the 129th annual meeting of the American Economic Association, Shiller (2017) stated that “as research methods advance, and as more social media data accumulate, textual analysis will be a stronger field in economics in coming years.” We aim to advance the literature in this exact direction.

There are three main approaches for measuring investor sentiment and/or disagreement in the economics literature. The first is to proxy sentiment using market-based measures such as trading volume, closed-end fund discounts, and initial public offering first-day returns, among others. Arguably, the most influential measure is Baker and Wurgler’s (2006) investor sentiment index, which is constructed as the principal component of six market-based proxies. The second approach is based on surveys. Popular sentiment measures include the University of Michigan Consumer Sentiment Index and the UBS/GALLUP Index of Investor Optimism. The dispersion

\(^1\) Akerlof and Shiller (2010) provide an updated elaboration on this concept in various economic contexts.
of the survey of professional forecasters (SPF) has been used as a proxy for disagreement (Ilut and Schneider (2014), Bollerslev et al. (2018)).

The third approach, which we adopt here, relies on textual data. Under this approach, empiricists have constructed measures for investor sentiment and disagreement using a variety of textual data, including online message posts (Antweiler and Frank (2004)), newspapers (Tetlock (2007), García (2013)), corporate 10-K reports (Loughran and McDonald (2011)), Google search records (Da et al. (2015)), and Federal Open Market Committee (FOMC) statements (Bollerslev et al. (2018)). This burgeoning literature is fueled by increasingly available computational tools for data gathering and natural language processing. Textual analysis has also been fruitfully used in other areas of economics: Gentzkow and Shapiro’s (2010) study on media slant and Baker et al.’s (2016) work on economic policy uncertainty are two recent influential examples, and Loughran and McDonald (2016) and Gentzkow et al. (2019) provide recent reviews on the broad range of textual analysis applications in accounting, economics, and finance.

The textual approach is complementary to more traditional market-based and survey-based approaches. Compared with market-based proxies, textual measures are “more primitive” in the sense that they do not directly rely on equilibrium market quantities (e.g., return or volume), which may be confounded by a plurality of market factors. Compared with survey-based proxies, textual measures are often available at higher frequency, whereas surveys, if available, are typically conducted monthly or quarterly. But the textual approach also has its limitations. An obvious drawback is that textual datasets are not readily available from standard databases. Gathering these massive datasets is costly, and the subsequent textual analysis often requires computational tools that might be foreign to applied economists, given the current state of the literature. The sense of “foreignness” can be literal when non-English textual data are involved (like in this study), as textual analysis in the mainstream economics literature has mostly focused on the English language.

Set against this background, we make three contributions. The first concerns data construction. We build a unique—and massive—textual dataset consisting of 60 million messages (or 6.6 billion Chinese characters) posted on a leading online investor forum in China spanning a 10-year sample period from 2008 to 2018. We manually construct a dictionary that collects words with positive and negative (i.e., optimistic and pessimistic) textual tones, which is customized to our empirical context. In addition, we manually label a subset of 40,000 messages for training machine-learning-based textual analysis methods. These efforts allow us to effectively use both the conventional dictionary method (Tetlock (2007), Loughran and McDonald (2011)) and supervised machine-learning methods (Vapnik (1995), Goodfellow et al. (2016), Trevor et al. (2017)) to extract textual
sentiment. The message-board dataset used in the pioneering work of Antweiler and Frank (2004) is the most similar to ours in the economics and finance literature. Their dataset consists of about 1.6 million messages from Yahoo! Finance and Raging Bull during the year 2000 (at the peak of the dot-com bubble), of which 1,000 messages are labeled for training their Naive Bayes algorithm. In contrast, our dataset contains significantly more observations and, importantly, spans a much longer and more representative time period, which is crucial for conducting reliable time-series econometric inference regarding aggregated stock market activities.

The second contribution pertains to measurement. We employ a broad variety of textual analysis tools to quantify the tone of the text messages. On one hand, we adopt the conventional dictionary-based method, which relies on the counts of positive and negative words specified by the customized dictionary described above. On the other hand, we employ state-of-the-art machine-learning methods for information extraction, including support vector machines (SVM) and convolutional neural networks (CNN). SVM is a popular machine-learning algorithm (Vapnik (1995), Trevor et al. (2017)), and has been applied in recent work on text-based empirical finance (see, e.g., Manela and Moreira (2017)). CNN is one of the so-called “deep-learning” methods that have been widely applied in many real-world applications.\(^2\) To the best of our knowledge, however, our application of CNN is the first in the context of measuring sentiment and disagreement in stock markets. In addition to these generic machine-learning methods, we also use specialized tools recently developed in computational linguistics. More specifically, we apply a representation method known as the word2vec (Mikolov et al. (2013)), which transforms words (in natural language) into numerical vectors so that semantically close words correspond to numerically similar vectors. As such, the words’ semantic meanings are partially preserved in their numerical representations, which, as we show, results in notable improvement in the effectiveness of machine-learning algorithms in our empirical context.

Using a standard out-of-sample evaluation scheme, we find that SVM and CNN, combined with word2vec representation, have the highest accuracy for predicting the sentiment labels given by human readers. But the performance of the dictionary method using our customized dictionary is also adequate. Market-level textual sentiment (resp. disagreement) indexes constructed using different types of textual analysis methods are highly correlated, suggesting that they are all alternative proxies for the same underlying sentiment (resp. disagreement) factor, and the measurement is robust to the choice of textual analysis method. Hence, we average the dictionary-, SVM- and CNN-based sentiment indexes into a single sentiment index, which we refer to as the China Investor

---

\(^2\)See, for example, LeCun et al. (2015) and Goodfellow et al. (2016).
As the third contribution, which is also our main economic contribution, we use these textual sentiment and disagreement measures to test a battery of predictions from classical behavioral asset-pricing models regarding the mispricing, underreaction, overreaction of stock prices to investor sentiment, excess volatility, and trading volume. Compared with previous literature, we test a broad set of economic hypotheses in a unified empirical setting, which is made possible by our unique textual dataset. As detailed below, our analysis uncovers many new empirical findings that support these theoretical predictions. Of course, another distinctive feature of this study is that we focus on China’s stock market using Chinese textual data. In so doing, we aim to diversify the textual analysis literature in economics and finance, which heavily concentrates on the English-speaking world, especially the U.S. market. Given the size of China’s economy, we believe it is of great economic importance to examine the predictions of classical behavioral finance theories in the enormous stock market of the world’s largest developing country. For readers who are not interested in the Chinese stock market per se, our study may be viewed as an out-of-the-sample check for previous empirical findings based on U.S. data.

Our analysis provides strong empirical support for a range of theoretical predictions. First and foremost, we find that textual sentiment can significantly predict market returns (De Long et al. (1990)), with its impulse response exhibiting both short-run underreaction and long-run reversal, as predicted by the theories of Barberis et al. (1998), Daniel et al. (1998), and Hong and Stein (1999). This evidence is consistent with the autocorrelation patterns of various aggregate stock index returns documented by Cutler et al. (1991), and is more generally related to cross-sectional evidence on momentum (Jegadeesh and Titman (1993)) and reversal (De Bondt and Thaler (1985)) for individual stock returns. We also find that the effect of textual sentiment on stock price is stronger and longer lasting for small and growth stocks than it is for big and value stocks, which is consistent with the limits-to-arbitrage argument, that is, the former are more difficult to arbitrage, and thus more sentiment-prone than the latter (Shleifer and Vishny (1997), Baker and Wurgler (2006)). Our findings can be contrasted with those of Antweiler and Frank (2004), who also used message-board data but did not find predictive power of their textual bullishness measure for individual stock returns. This might be due to their relatively short one-year sample, which also occurred during the peak of the dot-com bubble; the short sample span also rules out the possibility of studying longer-run effects concerning underreaction and overreaction.

---

3 We publish a summary of market- and industry-level CISIs every month in the China Securities Journal, the leading newspaper for China’s security markets. This paper is the first based on our textual sentiment indexes.

4 See Allen et al. (2005), Allen et al. (2009), Brunnermeier et al. (2017), Bian et al. (2018), Song and Xiong (2018), and Liu et al. (2019) for more detailed discussions on China’s financial markets.
Similar to our findings, Tetlock (2007) demonstrates, in his seminal paper, that a textual pessimism measure extracted from the “Abreast of the Market” column in the *Wall Street Journal* significantly predicts future stock returns in the U.S. market. The estimated effect is dispersed throughout the next trading day and is then reversed completely within a week. But it is important to note that this “fast” reversal provides no support for the well-known momentum and reversal phenomena that occur on the time scale of at least several months: These empirical regularities have been documented by De Bondt and Thaler (1985), Cutler et al. (1991) and Jegadeesh and Titman (1993), which in turn motivated the theoretical developments of Barberis et al. (1998), Daniel et al. (1998), and Hong and Stein (1999), among others. In sharp contrast to Tetlock’s (2007) finding, we document an underreaction-overreaction pattern on a longer time scale that is similar to these classical empirical and theoretical results, suggesting that our textual sentiment extracted from the message board indeed captures the noise trader sentiment considered in prior studies.

The difference between our findings and those of Tetlock (2007) suggests that the textual sentiment measures extracted from different media sources reflect different types of sentiment. The “Abreast of the Market” column mainly reflects the sentiment of the journalist (who may form his/her opinion by interviewing traders). Although the column may further influence its readers, the degree of effectiveness may change from time to time, and is not reflected in Tetlock’s textual pessimism measure. On the other hand, the textual sentiment extracted from our message data more directly reflects the opinions of a large number of individual investors, and hence is likely to be more representative of the noise trader sentiment that plays a key role in behavioral asset-pricing models. Our results thus support Shiller’s (2016) view (see Chapter 10) that conventional media, such as newspapers or television, have limited ability to activate investor behavior compared with face-to-face or word-of-mouth communications. In recent times, the latter types of interpersonal and interactive communications take place more frequently on social media, such as the online investor forum studied here. Shiller suggests that “these new and effective media for interactive (if not face-to-face) communication may have the effect of expanding yet again the interpersonal contagion of ideas.” The resulting herd behavior can amplify the aggregate effect of noise trading, and hence leads to more pronounced mispricing, underreaction, and overreaction patterns in the price dynamics.

Going one step further, we show empirically that the effect of textual sentiment on future stock returns is time-varying, with a stronger effect under higher investor attention and during more volatile periods. This finding is consistent with the theory of limited attention (Kahneman (1973)),
which suggests that information from the (broadly defined) media would not influence the stock price unless investors actually pay attention to it. In this analysis, we measure investor attention using the message-posting activity level and find that it dominates the more traditional attention proxy based on trading volume (Gervais et al. (2001), Barber and Odean (2008)). Meanwhile, our evidence for volatility-driven time variation also provides further support for Shleifer and Vishny’s (1997) limits-to-arbitrage theory in that, as argued by these authors, risk-averse arbitrageurs are less likely to trade against noise traders during volatile periods, resulting in an “untamed” impact from sentiment-driven noise trading. Interestingly, we find that attention and volatility can capture nearly all time variations uncovered by an alternative nonparametric rolling-window estimator. These findings are new to the literature, to the best of our knowledge. However, unlike García (2013), we find little support for the hypothesis that the effect of sentiment is stronger during economic downturns (measured by real GDP growth).

In addition to the return predictability, we also show that unusually high or low textual sentiment predicts high volatility and trading volume on the next day. These findings support De Long et al.’s (1990) prediction that noise trading induces excess volatility, as well as the classical notion of “overtrading” (Kindleberger and Aliber (2005)). Consistent with the voluminous literature on volatility modeling (Engle (1982), Bollerslev (1986), Nelson (1991), Engle and Ng (1993)), our estimates also reveal an asymmetric impact curve of sentiment on volatility, for which bearish sentiment exerts a larger influence on volatility than bullish sentiment.

Last but not least, we document that the next day’s trading volume is higher when there are more differences-of-opinion among investors, as measured by our textual disagreement index. This evidence supports the central prediction of a large class of disagreement models; see Harrison and Kreps (1978), Harris and Raviv (1993), Kandel and Pearson (1995), Scheinkman and Xiong (2003), Hong and Stein (2007), and references therein. Although this finding is well expected from theory, it is nevertheless empirically remarkable because, in prior work, Antweiler and Frank (2004) report the opposite result: Higher disagreement predicts lower trading volume in their analysis of message-board data from the year 2000 (while also documenting a positive contemporaneous correlation between disagreement and volume). To the best of our knowledge, our study is the first to document the positive predictive power of textual disagreement on trading volume. Further investigation into whether Antweiler and Frank’s (2004) finding is specific to the dot-com bubble episode, or may hold more generally in a longer and more representative sample in the U.S. market, would be an interesting question for future research.

The rest of the paper is organized as follows. We describe the data in Section II. In Section
III, we discuss textual analysis methods for extracting textual sentiment and construct textual sentiment indexes for China’s stock market. Section IV studies hypotheses related to sentiment, mispricing, underreaction, overreaction, and excess volatility. In Section V, we construct the textual disagreement measure and use it to test hypotheses related to trading volume. Section VI concludes. The appendix contains implementation details for the textual analysis methods used in this paper.

II. Data

In this section, we describe the datasets used in our empirical analysis. Section II.A discusses the textual data obtained from an online investor forum. Section II.B discusses the financial data we rely on to test economic hypotheses.

II.A. Textual Message Data

We download messages posted between July 1, 2008 and February 14, 2018 from Eastmoney, a leading online investor forum for stocks listed on the two stock exchanges in mainland China, Shanghai and Shenzhen. Each message contains a unique identifier of the subject company, a title, content, and time stamp with 1-second granularity. The leading stock market index in China is the CSI 300 index (China Securities Index), the constituents of which account for about two-thirds of the total market value and one-third of trading volume in the A-share market. In order to construct the corresponding market-level sentiment index, we focus on messages related to the 300 constituent stocks in the CSI 300. We drop duplicate messages and, following the standard practice in textual analysis, remove non-text items such as encoded images, tables, and HTML tags. In the A-share market, retail investors hold about 20% of the market value, but contribute more than 80% of the trading volume. The main reason for this disparity is that many large companies are effectively state owned, with a majority of shares held by government agencies with minimal (if any) trading. According to Liu et al. (2019), individual investors own 88% of the market’s free-

---

5 According to a recent analyst report by iResearch in 2018, Eastmoney is the top financial website in China. The effective monthly viewing time is 78 million hours, or 45% in terms of market share, which is larger than the remaining nine of the top 10 firms combined.

6 A-shares are denominated in renminbi (the Chinese currency) and are traded primarily between local investors. International investors mainly trade B-shares in U.S. (resp. Hong Kong) dollars on the Shanghai (resp. Shenzhen) stock exchange.

7 These statistics are obtained from the 2018 Statistics Yearbook of Shanghai stock exchange (see page 535). A Forbes article, “Five Reasons Why Global Investors Should Be Investing In China A-Shares,” dated May 28, 2018, also reported that retail investors accounted for 86% of total market trading volume in 2016.

8 For example, the largest firm in the Chinese stock market is the Industrial and Commercial Bank of China. The Ministry of Finance and the (state-owned) Central Huijin Investment Corporation control 34.6% and 34.71% of the bank’s domestic shares (A-share and H-share), respectively, which have been virtually constant during the 2015–2018 period.
floating shares. Meanwhile, a representative poster on the message board is very likely to be a retail investor. We thus expect that the textual information extracted from these messages adequately captures the investor sentiment that drives the majority of trading in China's stock market.

Table I reports summary statistics for the message data. The average number of messages each day is 17,029, and an average message contains 110 Chinese characters. From the table, we observe a heavily right-skewed distribution of message length. Unlike a typical (relatively short) message—in which the poster expresses his/her opinion in a few sentences—very long messages are often copied and pasted from other sources, such as news stories and analyst reports. We apply a simple filter to eliminate these potentially influential outliers, by retaining only messages with fewer than 500 Chinese characters. This criterion removes 4.85% of the messages from our sample, resulting in 59,875,650 messages for the subsequent analysis. Each message contains 47 characters on average.

Message posting activity exhibits strong intraday and within-week seasonalties. Figure I plots the average number of messages for each day of the week at half-hour frequency. As expected, weekdays have more messages than weekends, and posting activity is more intensive during trading hours than it is during non-trading hours. The total amount of messages is almost evenly split between trading and non-trading hours, however, with their shares being 49.5% and 50.5%, respectively.

Unlike Germanic and Latin languages, Chinese words in a sentence are not separated by spaces. In order to conduct our textual analysis, we parse the sentences into separate words using standard text-segmentation software. To reduce the vocabulary to a manageable size, we follow a common practice (see, e.g., Gentzkow et al. (2019)) to remove punctuation, numbers, “stop” words (i.e., very common words such as the Chinese equivalents of “this,” “that,” “and,” “or,” etc.), and very rare words that appear fewer than 10 times in the entire sample of 60 million messages. This parsing process identifies a vocabulary of 654,555 unique Chinese words in the sample.

The most commonly used approach for extracting textual sentiment is the dictionary method (see, e.g., Tetlock (2007), Loughran and McDonald (2011, 2016), and García (2013)).
approach, one measures the textual sentiment of a message using the counts of positive and negative words specified by the corresponding word lists in a dictionary. As emphasized by Loughran and McDonald (2011, 2016), the effectiveness of this approach depends crucially on the relevance of the dictionary with respect to the empirical context at hand; this requires a nontrivial amount of domain knowledge and human input.

We consider two dictionaries in our study. The first (used as a benchmark) is the Chinese translation of the financial dictionary created by Loughran and McDonald (2011), henceforth LM. The LM dictionary collects positive and negative (English) words, in a financial sense, from corporate 10-K reports. We use three translation tools to translate the 353 positive words and 2,337 negative words in the LM dictionary into Chinese. Each English word is translated into several Chinese words, and we manually screen the translation to remove unintended ones. This process results in a Chinese version of the LM dictionary with 608 positive words and 2,274 negative words. We refer to this dictionary as the **Chinese LM dictionary**.

Given the many successful applications of the LM dictionary in the literature, we naturally expected that its translated version would also perform well in our application. However, a shortcoming of the Chinese LM dictionary soon became evident in our subsequent analysis of message data: The messages often contain jargon and informal language (i.e., slang), which are highly informative about textual sentiment, but are not included in the Chinese LM dictionary. A further manual inspection reveals that this issue is not due simply to noise in the translation process, because slang would seem foreign even to an average native Chinese speaker. This is hardly surprising, because the original (English) LM dictionary is constructed from formal corporate reports, and the semantic formality is intentionally preserved in our translation. The language used in the message data, in contrast, is generally far more casual.

A human reader’s input is thus needed to remedy this issue in order to effectively implement the dictionary method. At the same time, machine-learning algorithms also demand a training set of messages labeled by human readers. As an important component of our data construction, we employed a team of research assistants (economics graduate students at Peking University) to read a random subsample of 40,000 messages. These readers manually labeled each message as one of the three types, {positive, negative, ambiguous}, and provided a list of keywords that informed their judgment. For example, a message is labeled as “positive” or “negative” if the poster recommends

---

11The translation tools include Google Translation and its Chinese analogues, Baidu Translation and Youdao Translation.

12For example, the Chinese word “Hong-San-Bing” (in romanization), literally translated as “three red soldiers,” indicates a three-day consecutive run-up of the stock price after a period of drops, and is used as a signal for bullish reversal by technical analysts. (The red color indicates a price increase in China’s stock market.)
that others buy or sell the stock; some messages (e.g., advertisements) do not have a clear positive or negative tone, and hence are labeled “ambiguous.”

To streamline the discussion, we describe the exact protocol of the labeling process in Appendix I, but highlight a few key features here. First, a 4-hour training session is provided to readers to reduce judgmental idiosyncrasies among them. Second, readers must justify their judgments by further providing keywords on which each decision is based. Third, each reader’s pecuniary reward depends on the quality of his/her output, as determined by a randomized audit. Finally, each message is read by two readers, and we consider their labels valid only if they agree. Of the 40,000 messages, we obtain 27,999 valid labeled messages (9,288 positive, 15,373 negative, 3,338 ambiguous), together with a list of keywords. Finally, we manually audited 2,500 of the labeled messages and reviewed all keywords, and deemed the outcome of the labeling process to be satisfactory.

For the subsequent textual analysis, we augment the Chinese LM dictionary with these keywords. The resulting dictionary will be referred to as the Chinese Stock Market (CSM) dictionary. Labeled messages will be used to train machine-learning algorithms for textual analysis, as discussed in Section III.

II.B. Financial Data

Financial data are collected from three databases: WIND, RESSET, and CSMAR. From WIND, we collect (1) daily CSI 300 price and trading volume series; (2) daily indexes for international asset markets, including the S&P 500 index for the U.S. stock market and the NYMEX oil futures; and (3) tick-level price data from 2016 to 2018. Tick-level data from earlier periods (2008–2015) are collected from RESSET. These high-frequency data are used to calculate intraday returns and daily realized volatilities (see, e.g., Andersen et al. (2003)). All firm characteristics data are taken from CSMAR, including the CSI 300 constituent stock list, which is typically adjusted semiannually.

\[ \text{[INSERT FIGURE II]} \]

---

13 For example, if one reader labels a message as positive and the other labels the same message as negative, we consider the labeling of this message invalid. A seemingly reasonable alternative is to put this message into the ambiguous category. But the latter protocol is prone to the readers’ potential shirking behavior. For example, consider a message that is clearly positive. A diligent reader will label it as positive, but the “shirking” reader may randomly label it as negative. Putting this message into the ambiguous category would be a mistake and contaminate the quality of our dictionary and the training set. Our protocol is designed to mitigate this issue, at the cost of losing some observations in this (costly) labeling process; but this choice serves our long-term goal of gradually building a larger high-quality training dataset. Formally addressing readers’ potential moral hazard issue is clearly beyond the scope of the current paper.
Figure II plots the daily closing price and trading volume (in shares) of the CSI 300 index. The most notable episode in our sample is the stock market run-up in 2015, followed by the largest crash over the past 10 years.

III. TEXTUAL SENTIMENT INDEXES

In this section, we describe the construction of the China Investor Sentiment Index (CISI). Section III.A discusses the textual analysis methods used to extract textual tone from individual messages. In Section III.B, we construct the market-level textual sentiment indexes and report their basic properties.

III.A. Extracting Textual Sentiment from Message Data

In this subsection, we describe the methods we rely on to extract textual sentiment from the messages. As discussed in the data section, we classify the textual sentiment of a message into one of the three categories, \{positive, negative, ambiguous\}. Ideally, with an unlimited amount of input from a representative human reader, we would hope to classify all text messages; the aggregate textual sentiment of these messages can be measured by the proportions of positive and negative ones (we consider ambiguous messages to be uninformative). In practice, manually classifying the tens of millions of text messages in our sample would obviously be very costly. We instead rely on computational algorithms that mimic the representative human reader’s classification. We consider three types of methods: dictionary, support vector machine (SVM) and convolutional neural network (CNN).

The dictionary method is perhaps the most commonly used approach in the textual analysis literature of accounting, economics, and finance; see, for example, Tetlock (2007), Loughran and McDonald (2011, 2016) and García (2013). In our setting, a dictionary contains a list of positive words and a list of negative words, which we denote PosList and NegList, respectively. The textual sentiment of each message is determined by its word counts in these word lists. Let $W_{i,j}$ denote the number of times that word $j$ appears in message $i$. Following Loughran and McDonald (2011), we further transform the raw word count into the following weight

$$
\tilde{W}_{i,j} = \frac{1 + \log(W_{i,j})}{1 + \log(\text{number of words in message } i)} \times \log \left( \frac{\text{total number of messages}}{\text{total number of messages that contain word } j} \right),
$$

where the first term adjusts for the document’s length and the second term adjusts for the word’s commonality across the entire sample. The sentiment score of message $i$ is then defined as the
relative weight of positive and negative words:

\[
\text{score}_i = \frac{\sum_{j \in \text{PosList}} \tilde{W}_{i,j} - \sum_{j \in \text{NegList}} \tilde{W}_{i,j}}{\sum_{j \in \text{PosList}} \tilde{W}_{i,j} + \sum_{j \in \text{NegList}} \tilde{W}_{i,j}},
\]

(dictionary method).

A higher sentiment score indicates a more positive tone of the message. We compute the sentiment score using both the Chinese LM dictionary and the CSM dictionary.

SVM and CNN are popular machine-learning methods that have been widely used in many academic and commercial areas.\(^{14}\) For our purposes, these machine-learning methods play essentially the same role as the multinomial logit model commonly used in econometrics to model categorical data: We estimate model parameters by treating the labels provided by the human readers as the dependent variable and using the text messages as the explanatory variable.\(^{15}\) The estimated model can then be used to compute the probability \(\text{prob}_{i,c}\) that each message \(i\) belongs to category \(c\), \(c \in \{\text{positive, negative, ambiguous}\}\), for all messages including the unlabeled ones. The sentiment score of message \(i\) is then defined as

\[
\text{score}_i = \text{prob}_{i,\text{positive}} - \text{prob}_{i,\text{negative}},
\]

(SVM or CNN method).

We describe the computational details in Appendix II, which also includes a comparison of the classification accuracy of the dictionary, SVM, and CNN methods (together with some additional variants) in a standard out-of-sample evaluation framework. We find that SVM and CNN have the highest, and very similar, classification accuracies (81%) for mimicking a human reader’s judgment. The performance of the dictionary method depends crucially on which dictionary is used: Our customized CSM dictionary attains 74% accuracy, which is adequate in comparison with the machine-learning methods and much higher than the 33% accuracy obtained using the Chinese LM dictionary. The latter finding is not surprising, because the original LM dictionary was constructed for a very different context, both linguistically and economically. Our result actually echoes Loughran and McDonald’s (2011) original insight, that it is important to tailor the dictionary to the specific empirical context in order for the dictionary method to be effective.

We stress that our intention in conducting the above comparison is not to find the “best” method; instead, we aim to rule out methods that are clearly underperforming. After all, classification

---

\(^{14}\)See Vapnik (1995), James et al. (2013), and Trevor et al. (2017) for additional discussion of the SVM method and Manela and Moreira (2017) for a recent textual application in finance. For the CNN method, see LeCun et al. (2015), LeCun et al. (1998), Kim (2014), Goodfellow et al. (2016), and the references therein.

\(^{15}\)An important component of implementing these machine learning methods is how to represent the “input,” originally in the form of raw text, as a numeric vector. Following the recent literature in computational linguistics (Mikolov et al. (2013), Mikolov et al. (2013)), we use the \textit{word2vec} method, which represents semantically close words with likewise close numeric vectors; see Appendix II.B for details.
accuracy only reflects one specific aspect of the sentiment extraction, whereas the importance of simplicity and transparency has been emphasized by Loughran and McDonald (2016), who caution that “more complex methods potentially add more noise than signal.” Machine-learning methods, especially deep-learning algorithms such as CNN, are perhaps still deemed by many economists to be a “black box,” while the dictionary method is easier to understand. That being said, the notion of transparency may not be as clear as one might think. Consider this study as its own example. The CSM dictionary constructed in this paper is in Chinese, and contains extensive stock market jargon and slang. While an experienced Chinese investor can easily interpret the dictionary, a reader who does not understand Chinese may find our dictionary far less transparent. In this context, the reader may find the machine-learning methods to be more transparent because the algorithms are trained in essentially the same way regardless of the specific language. From this perspective, the generic nature of machine-learning methods improves transparency, and hence renders these methods (at least) useful complements to the traditional dictionary method. For this reason, we use all three types of methods (CSM dictionary, SVM, and CNN) to construct textual sentiment indexes for our economic analyses.

**III.B. Textual Sentiment Indexes for China’s Stock Market**

We proceed to construct market-level textual sentiment indexes by aggregating the sentiment scores of individual messages. We use all three types of methods (CSM dictionary, SVM, and CNN) to generate three versions of sentiment scores for each message. We remind the reader that the sentiment score takes (continuous) values in the $[-1, 1]$ interval, and a higher value indicates a more optimistic tone.

For each stock $i$ in the CSI 300 index, we construct the daily textual sentiment index on trading day $t$ using the average sentiment scores of all messages posted for this stock during the period between 3:00 pm on trading day $t - 1$ and 3:00 pm on trading day $t$. This timing convention matches that of close-to-close daily stock returns. (In later analyses, we also consider an overnight sentiment index formed using messages during the close-to-open period; see Section IV.) The market-level textual sentiment index is then computed as the value-weighted average of stock-level sentiment indexes, and we also construct an equally weighted version for robustness checks.\(^{16}\) This aggregation process is done separately for each of the three textual analysis methods, resulting in three versions of market-level textual sentiment indexes. We refer to these versions as the Dictionary, SVM, and

\(^{16}\)The value-weighted index may put too much weights on big stocks, in that a majority of their shares are held by the government through various agencies. For these stocks, the free-floating part of the market value is far less than the total market capitalization. The equally weighted indexes provide robustness checks regarding this issue.
CNN sentiment indexes.

[INSERT TABLE II]

Since these indexes are formed using distinctly different textual analysis methods, no mechanical reason guarantees their similarity. However, if they are indeed proxies for the same underlying textual tone, their correlations should be high. As a sanity check, we report the correlations between the Dictionary, SVM, and CNN sentiment indexes in Table II. For value-weighted indexes, the correlation between SVM and CNN sentiment indexes is 0.96, and their correlations with the Dictionary sentiment index are 0.90 and 0.87, respectively. The equally weighted indexes show essentially the same pattern. These high correlations suggest that the different indexes measure the same latent “true” textual sentiment and our measurement is robust to the choice of the specific textual analysis method. To simplify our discussions in the empirical analysis presented in later sections, we further aggregate the Dictionary, SVM, and CNN indexes by taking a simple average of them, which we refer to as the China Investors Sentiment Index (CISI). By construction, the CISI is highly correlated with each of the three individual components. We report our findings mainly based on the CISI, noting that the results are qualitatively unaltered when we use its individual components instead. All index series are standardized to have zero mean and unit standard deviation.

[INSERT FIGURE III]

[INSERT TABLE III]

Figure III plots the daily time series of the (value-weighted) CISI, along with its 22-day moving average. The CISI series appears to be stationary and exhibits a nontrivial degree of persistence. We also see the large drop in the textual sentiment during the 2015 market crash. Table III provides summary statistics for the daily CISI and market activity variables. The distribution of the CISI is close to the normal distribution, with nearly zero skewness (-0.02) and mild excess kurtosis (0.127). In the fifth row, we report the correlation between the CISI and the contemporaneous daily stock return, log volume, and log realized volatility, where the realized volatility is computed as the sum of squared 5-minute returns within the regular trading hours (Andersen et al. (2003)). As expected, the CISI is positively correlated with stock return (correlation = 0.24). We also see that higher trading volume and volatility are associated with lower textual sentiment, although the magnitudes of these correlations are moderate. The autocorrelation coefficients of the daily CISI series are sizable: The first-order autocorrelation is 0.643, and drops by roughly half at the 22nd
lag (i.e., 1 trading month). Although the textual sentiment is clearly related to stock returns, these series exhibit drastically different dynamics, as the daily stock returns are essentially uncorrelated. Compared with the C ISI, the volume and the realized volatility series are more persistent, as is evident from their high and slowly decaying autocorrelations.

The correlation and autocorrelation statistics in Table III suggest that the daily C ISI is related to, but distinct from, each of the market activity variables. As a way of quantifying the contemporaneous association between the C ISI and these market variables jointly, we regress the day- \( t \) \( C ISI \) on the same-day stock return \( (R_t) \), log volume \( (\log Vlm_t) \), and log realized volatility \( (\log RV_t) \), while controlling for weekday dummies and lagged return, log volume, and log realized volatility variables up to 5 days. The estimated model is

\[
CISI_t = 9.068 + 0.138 \times R_t - 0.084 \times Vlm_t - 0.005 \times RV_t + \hat{\delta} \times Controls_t + \hat{\epsilon}_t,
\]

where Newey–West standard errors computed with eight lags are reported in parentheses. This regression further confirms that the positive (conditional) association between the C ISI and stock return is statistically significant. That said, the adjusted R-square of this regression is 33.5%, suggesting that a large part of the temporal variation in the C ISI cannot be captured by contemporaneous and lagged market variables.

IV. Textual sentiment and stock returns

In this section, we test several hypotheses regarding textual sentiment, stock return, and volatility. Section IV.A presents baseline results for the Granger causality of textual sentiment on stock price. Further results on underreaction and overreaction, cross-sectional heterogeneity, and time-varying effects are presented in Sections IV.B–IV.D. Section IV.E discusses the impact of textual sentiment on volatility.

IV.A. Granger Causality of Textual Sentiment on Stock Price

Textual sentiment may be informative about stock price movements for two theoretical reasons. First, it may contain fundamental information that is not yet incorporated into the price. Conventional asset pricing theory suggests that new fundamental information should cause the price to change. Second, it may capture investors’ sentiment in the sense of De Long et al. (1990), or Keynes’s notion of animal spirit (Keynes (1936), Akerlof and Shiller (2010)). A positive tone in the text message reflects a bullish sentiment that in turn exerts upward price pressure, and vice versa.
The corresponding econometric analysis is complicated by the issue of dynamic feedback first articulated by Granger (1969): In the present context, as asset prices fluctuate during the trading day, price information is gradually integrated into text messages. To the extent that textual sentiment is “continuously intertwined” with the price information within the trading day, a contemporaneous regression for these daily series is difficult to interpret economically. As a result, prior research often focuses on whether textual sentiment can predict stock prices and other market variables, which amounts to studying the Granger causality of textual sentiment on future stock market activities.

Empirical findings are mixed. Tetlock (2007) shows that the textual pessimism embedded in the “Abreast of the Market” column in the *Wall Street Journal* statistically significantly predicts negative returns the next day. García (2013) reports similar findings using information extracted from two columns in the *New York Times*. Loughran and McDonald (2011) show that a more negative tone in a corporate 10-K report precedes lower stock returns. Note that these studies rely on textual information reported by journalists or firm executives, which is distinctively different from the message-board data we use in this paper. By their very nature, newspaper articles and 10-K reports are mostly factual “snapshots” of the market and the companies, and they do not directly reflect the opinions of individual investors. Antweiler and Frank (2004) use message-board data similar to ours from Yahoo! Finance and Raging Bull for the year 2000, but they find that their textual sentiment (i.e., “bullishness”) does not have significant predictive power for stock prices.

Set against this background, in this subsection we examine the Granger causality of our textual sentiment indexes on stock prices. As emphasized by Granger (1969), the aforementioned “feedback” issue is more severe when the data available to the econometrician are updated more slowly. Following his insight, we mitigate this issue by exploiting a unique feature of our textual dataset: Since the high-frequency time stamps of the message data are available, we can concentrate on the textual information accumulated “overnight” between the close of the previous trading day and the opening of the new trading day. Since post-market price discovery is very limited in China’s stock market, overnight textual information is much less influenced by concurrent market price signals (hence less feedback). We then examine whether overnight textual sentiment can predict the next day’s stock price.

To be precise, we illustrate in Figure IV the timeline underlying this predictive analysis. The

---

new trading day, $t + 1$, will begin with a 10-minute call-auction session at 9:15 am Beijing time, and after a short break the continuous-auction session will start at 9:30 am. Before the market opens, the econometrician has formed a conditioning information set containing various sources of information, as illustrated in Figure IV. This includes lagged market information (price, trading volume, volatility) up to 3:00 pm (market close) on day $t$. Afterward, the econometrician also reads the messages posted during the overnight period from 3:00 pm to 9:15 am, and summarizes them by aggregating their textual tones into an overnight textual sentiment index. Although the local stock market is closed during the overnight period, the econometrician can gather price information from international asset markets, such as the S&P 500 index in the U.S. stock market and the price of oil futures, which are also closely monitored by Chinese investors. Equipped with these sources of conditioning information, the econometrician aims to predict the 9:30 am opening price; evidently, since the closing price on day $t$ is already in the conditioning information set, predicting the opening price is equivalent to predicting the overnight return, defined as

$$R_{o,t+1} = P_{9:30 \text{ am day } t+1} - P_{3:00 \text{ pm day } t},$$

where $P_{\tau}$ denotes the logarithm of the CSI 300 index at time $\tau$.

We examine the Granger causality of overnight textual sentiment on overnight returns using the following specification (similar regressions with longer predictive horizons up to 1 year are presented in Section IV.B),

$$R_{o,t+1} = \alpha + \beta \times S_{o,t} + \delta \times Controls_t + \epsilon_{t+1},$$

where the overnight textual sentiment index $S_{o,t}$ is constructed similar to the daily index, except that we only use messages posted during the overnight period.\(^{18}\) Controls$_t$ variables include the lagged values of daily stock return, log volume, and log realized volatility in the local stock market up to 5 days; the latest daily close-to-close returns of the S&P 500 index ($R_{US,t}$) and the NYMEX oil futures ($R_{Oil,t}$); and weekday dummies.\(^{19}\) As depicted in Figure IV, all conditioning variables on the right-hand side of equation (4) are known before the market opens on trading day $t + 1$.

\[\text{INSERT TABLE IV}\]

\(^{18}\)Similar to our construction of the daily sentiment index in Section III.B, we construct three versions of the overnight sentiment index using the Dictionary, SVM, and CNN methods, which are then averaged into the overnight CISI. We consider both value-weighted and equally weighted aggregation schemes (across individual stocks) when constructing the corresponding market-level indexes.

\(^{19}\)In results not reported here, we also used the return of COMEX gold futures as a control variable. However, the regression coefficient of this variable is not significant, and its inclusion has minimal effects on the estimation.
We now turn to the empirical findings. To set the stage, we report in Panel A of Table IV the baseline estimation results for equation (4) without any control variables. For comparison, we also consider an alternative specification with the overnight textual sentiment $S_{o,t}$ replaced by the lagged daily textual sentiment $S_t$, and report the corresponding estimation results in Panel B; recall that the $S_t$ index is formed using messages posted between 3:00 pm on day $t - 1$ and 3:00 pm on day $t$. For completeness, we present these baseline results for indexes generated by all four methods (Dictionary, SVM, CNN, and the averaged CISI) and for both value-weighted and equally weighted cross-sectional aggregation schemes.

Looking at Panel A1 of Table IV, we see that value-weighted overnight textual sentiment can strongly predict the overnight return. The estimated coefficients of overnight sentiment indexes generated by different methods are all positive and significant at the 1% level, and have similar magnitudes. These estimates are also economically large. For example, a one-standard-deviation increase in the overnight CISI predicts a 19.3-basis-point increase in the overnight return. These findings are qualitatively unaltered, and quantitatively slightly more pronounced, when we instead use equally weighted overnight textual sentiment indexes, as shown in Panel A2. Interestingly, the “catch-all” CISI measure is associated with the highest $R^2$, suggesting that it is a more accurate measure of underlying textual sentiment.

The estimation results are drastically different when we replace overnight textual sentiment $S_{o,t}$ with the lagged daily sentiment $S_t$. As shown in Panel B of Table IV, the coefficients of $S_t$ are close to zero and generally statistically insignificant, accompanied by adjusted $R^2$-squares that are virtually zero. This sharp contrast clearly highlights the advantage of overnight textual sentiment relative to the (more stale) lagged daily sentiment in these predictive regressions. Hence we focus exclusively on overnight textual sentiment in subsequent analyses.

[INSERT TABLE V]

The predictive power of overnight textual sentiment may be due to its correlation with other variables. To better understand this issue, we report in Table V the estimation results for equation (4) with various sets of control variables. For brevity, we only present results for the CISI. First, we observe from column (1) that controlling for the lagged stock return, volume, and volatility results in little change in the sentiment coefficient. Hence, the predictive power of the overnight CISI is not merely due to its correlation with lagged market information. Second, column (2) shows that the coefficient of the S&P 500 return is significantly positive and economically large. Including this variable also greatly improves the $R^2$. This is not surprising, because the U.S. stock market is closely watched by investors around the world. Since the $S_{o,t}$ variable is positively correlated
with the S&P 500 return (correlation = 0.188), its coefficient becomes smaller when the latter is controlled for, indicating that overnight textual sentiment picks up information from the U.S. stock market. That being said, the estimated effect of the CISI remains highly significant, both statistically and economically. The specification in column (3) further controls for lagged return, volume, and volatility variables up to 5 days. However, these additional control variables have little impact on the sentiment coefficient. Finally, we note that the results for the equally weighted CISI reported in columns (4)–(6) are essentially the same as those for the value-weighted CISI reported in the first three columns. Overall, these estimation results show that overnight textual sentiment can indeed predict the opening price on the upcoming trading day, above and beyond other more easily quantifiable variables from local and international asset markets.

IV. B. Underreaction and Overreaction

A further question is to what extent textual sentiment captures fundamental information and investor sentiment. Both theories clearly have an element of truth, and disentangling the two forces is difficult for at least two reasons. First, whether a message contains fundamental information not yet incorporated into the price cannot be determined by simply reading it, because an article that discusses firm fundamentals may be “tag-along news” (Shiller (2016)) that contains no new information at all. Black (1986) suggests that seemingly fundamental news may actually be noise. Second, fundamental information and investor sentiment are tied together, even in theory: Aggregate investor sentiment manifests as investors’ less than rational perception of the news they observe, which potentially results from conservatism and representativeness bias (Barberis et al. (1998)), overconfidence and self-attribution (Daniel et al. (1998)), or gradual information diffusion and myopic trend-chasing behavior (Hong and Stein (1999)).

We can nevertheless shed some light on this question by examining the impulse response of stock prices with respect to textual sentiment. The idea is straightforward: If textual sentiment mainly carries fundamental information, its effect on stock price should be realized quickly and extend to long horizons (i.e., an immediate and permanent effect); in contrast, to the extent that textual sentiment captures investors’ behavioral bias, the cumulative impulse response curve of stock price should exhibit short-run underreaction and long-run reversal, as predicted by the aforementioned behavioral asset-pricing theories. Specifically, Hong and Stein (1999) make explicit theoretical predictions for hump-shaped cumulative impulse response curves (see their Figures 1–4).

We adopt Jordà’s (2005) local projection method to estimate cumulative impulse responses.
Specifically, we consider the following multiple-horizon predictive regressions:

\[ P_{\tau} - P_{3:00 \text{ pm day } t} = \alpha_{\tau} + \beta_{\tau} \times S_{o,t} + \delta_{\tau} \times \text{Control}_t + \epsilon_{\tau}, \]

where \( \tau \) ranges from 3:00 pm on day \( t + 1 \) to 3:00 pm on day \( t + 250 \), extending the horizon from 1 day to 1 year. The multiple-horizon predictive regression is commonly used in similar contexts (see, e.g., Cutler et al. (1991)). The cumulative impulse response of the price with respect to \( S_{o,t} \) is captured by \( \beta_{\tau} \) coefficients at different horizons, which can be separately estimated by ordinary least squares. As shown by Jordà (2005), this local projection method is more robust to misspecification than the conventional vector-autoregression (VAR) for estimating impulse responses. We use the full set of control variables, as in column (3) of Table V, which includes the latest daily returns of the S&P 500 index and NYMEX oil futures, weekday dummies, and lagged daily returns, log volume, and log realized volatilities in the local market up to 5 days.

Table VI reports estimated sentiment coefficients at different horizons. We see that the cumulative impulse response gradually increases, peaks around 10 or 20 weeks (depending on the weighting scheme), and reverses to a lower and statistically insignificant level at longer horizons. Note that the sentiment coefficients for longer-horizon regressions are much larger than those for overnight returns, as shown in Table IV, suggesting that excluding the latter from the former would not change the pattern seen in Table VI. The salient pattern of short-run underreaction and long-run reversal is consistent with the predictions of standard behavioral asset-pricing models with less than rational traders; see De Long et al. (1990), Barberis et al. (1998), Daniel et al. (1998), and Hong and Stein (1999). The estimated hump-shaped impulse response pattern matches Hong and Stein’s (1999) theoretical prediction particularly well, and also mirrors Shiller’s (2017) general observations regarding social epidemics. Furthermore, this finding corroborates the short-run positive and long-run negative autocorrelation patterns documented by Cutler et al. (1991) for various asset market indexes.

This evidence suggests that our textual sentiment indexes indeed capture the investor sentiment theoreticized in classical behavioral asset-pricing models. Meanwhile, we do not find statistically significant evidence for a permanent price impact of textual sentiment, in that the post-reversal sentiment coefficients are all statistically insignificant. While it is reasonable to believe that textual sentiment may contain some fundamental information, the corresponding price impact (if any) appears to be statistically less salient than the impact exerted by the “animal spirit” component in the sentiment index.
In order to position these findings in a broader context, we compare them with prior empirical results for the U.S. market. Similar to our study, Antweiler and Frank (2004) construct a textual sentiment measure using messages from Yahoo! Finance and Raging Bull message boards. But these authors find that their textual sentiment measure does not have significant predictive power for stock returns. This lack of statistical significance may be due to their short (1-year) sample span, as it is well known that statistically powerful inference on expected returns demands a long sample period. The short sample span also prevents these authors from studying underreaction and overreaction at longer horizons, which is the focal point of the theoretical analysis of Barberis et al. (1998), Daniel et al. (1998), and Hong and Stein (1999), among others.

The underreaction and overreaction patterns documented in Table VI are also distinct from the findings of Tetlock (2007). Tetlock shows that his textual pessimism measure exerts a significantly negative impact that is dispersed throughout the next trading day, which is then reversed completely within 1 week. He thus concludes that the newspaper-based textual sentiment provides no support for theories of underreaction or overreaction at longer horizons (see p. 1150). Indeed, the within-week reversal is apparently “too fast” to be squared with the autocorrelation patterns of stock index returns documented by Cutler et al. (1991).20 This disparity suggests that the newspaper-based textual sentiment is unlikely to capture the type of investor sentiment that drives the underreaction and overreaction phenomena documented empirically by Cutler et al. (1991), Jegadeesh and Titman (1993), and De Bondt and Thaler (1985), which in turn motivated the theories of Barberis et al. (1998), Daniel et al. (1998), and Hong and Stein (1999). This is perhaps unsurprising, because the “Abreast of the Market” column used by Tetlock is the journalist’s summary of the market on the prior day, but the journalist’s sentiment (possibly influenced by his/her interviewees) may not adequately represent the aggregate sentiment of noise traders. In contrast, our textual sentiment indexes are directly extracted from a large number of messages posted by individual investors—who also contribute to the majority of the trading volume in the Chinese stock market—and the resulting impulse response shows a salient underreaction-overreaction pattern at horizons that are broadly consistent with the aforementioned empirical studies on momentum and reversal.

20Cutler et al. (1991) show that stock index returns exhibit momentum on the time scale of several months and show reversal at longer horizons. On the cross-section, Jegadeesh and Titman (1993) document momentum over 3- to 12-month holding periods, and De Bondt and Thaler (1985) document long-term reversal over 3- to 5-year holding periods.
IV.C. Cross-sectional Heterogeneity and Limits of Arbitrage

Our empirical evidence so far suggests that textual sentiment indexes capture investor sentiment and have predictive power for future stock returns. In this subsection, we provide further corroborative evidence guided by the limits-to-arbitrage theory. De Long et al. (1990) and Shleifer and Vishny (1997) point out that the price impact of sentiment-driven noise trading should depend on the extent to which arbitrage is limited. Hence, if textual sentiment indeed measures investor sentiment, its effect on harder-to-arbitrage stocks (e.g., small stocks) should be larger and longer-lasting.

We examine this hypothesis by considering stock portfolios formed according to the market capitalizations of individual firms. Specifically, we form two size-based portfolios that consist of the largest and the smallest 100 stocks of the constituents of the CSI 300 index, respectively; the portfolios are value-weighted and rebalanced monthly. Since small stocks are often considered to be more difficult to arbitrage, we expect the effect of textual sentiment to be stronger and longer lasting for the small-stock portfolio than the big-stock portfolio. In order to highlight the differential behaviors of these portfolios, we focus on the return of the small-minus-big (SMB) portfolio, constructed as the difference between returns of the small- and big-stock portfolios. We repeat the multi-horizon predictive regressions in equation (5) for the cumulative returns of the SMB portfolio, and report horizon-specific sentiment coefficients in Table VII, which measure the differences in the cumulative impulse responses of the two size-based portfolios.

From the table, we see that across all horizons, the sentiment coefficient for predicting the SMB portfolio return is positive, economically sizable, and generally statistically significant, suggesting that small stocks are indeed more sentiment-prone. Moreover, we also see that the sentiment coefficients are significant at longer horizons, which indicates that the effect of sentiment on small stocks is longer lasting than big stocks. This evident differential behavior between small and big stocks is particularly interesting when contrasted with the insignificant long-horizon sentiment coefficients for the market portfolio shown in Table VI, which suggests that this finding is not simply driven by the different loadings of the two size-based portfolios on the market factor. These results are, again, robust to our use of value-weighted or equally weighted sentiment indexes.

As an additional check, we sort stocks into a growth portfolio and a value portfolio according to their earnings-to-price ratios (E/P), which consist of the 100 stocks with the lowest and the
highest E/P ratios in the CSI 300 index, respectively. Since short selling is very costly in China, overpriced stocks are much more difficult to arbitrage. Hence, we expect that textual sentiment exerts a stronger and longer-lasting effect on growth stocks than it does on value stocks. Table VIII presents the cumulative impulse responses of the growth-minus-value portfolio with respect to textual sentiment. The findings are qualitatively similar to those for the SMB portfolio shown in Table VII, providing further support for the prediction from the limits-to-arbitrage theory.

**IV.D. State-dependent Effect of Textual Sentiment**

The effect of textual sentiment on stock return may be time-varying for at least three related reasons. The first pertains to investors’ attention (Kahneman (1973, 2011)). Textual information, regardless of whether it captures fundamental information or investor sentiment, would not affect asset prices unless investors actually pay attention to it. Therefore, the effect of textual sentiment on stock price should be higher when investors pay more attention to the message board, and vice versa. The dynamic aspect of this effect is described as an “attention cascade” by Shiller (2016). Second, Shleifer and Vishny (1997) argue that risk-averse arbitrageurs might avoid extremely volatile markets. We thus expect to see a larger effect of sentiment during more volatile episodes, because sentiment-driven noise trading is less absorbed, or corrected, by arbitrageurs in this situation. Finally, as argued by García (2013) on the basis of a large psychology literature, investors’ reaction to news will be more pronounced during periods of anxiety and fear; correspondingly, García (2013) documents that the effect of textual sentiment is stronger during economic downturns.

To proceed with the empirical investigation, we measure volatility using log realized volatility ($RV_t$) and use an indicator variable for below-median quarterly real GDP growth to signify economic downturns ($LowGrowth_t$). We also need measures for investors’ attention. Conventional proxies for investor attention include extreme returns (Barber and Odean (2008)) and trading volume (Gervais et al. (2001), Barber and Odean (2008), and Hou et al. (2008)). Da et al. (2011) argue that these equilibrium price or volume quantities measure attention only indirectly, because they may be driven by factors unrelated to investor attention. These authors instead propose an attention measure based on the Search Volume Index provided by Google Trend.

---

21Liu et al. (2019) suggest that the E/P ratio is a better alternative to the conventional book-to-market ratio for classifying growth and value stocks in China’s stock market. Our empirical findings remain qualitatively unaltered when we sort portfolios based on the book-to-market ratio instead.


23See, for example, Smith and Ellsworth (1985), Ortony et al. (1990), Tiedens and Linton (2001), Gino et al. (2012), and Akerlof and Shiller (2010).
We measure investor attention by closely mirroring prior work. Log trading volume ($Vlm_t$) is used as a conventional benchmark. In addition, we construct a more direct measure for investor attention by using the log number of messages posted during the overnight period, denoted $A_{o,t}$, which is then standardized to have zero mean and unit standard deviation. In the same spirit as the search-volume-based measure of Da et al. (2011), the $A_{o,t}$ variable is also a “revealed” attention measure in the sense that if the investor is willing to exert the effort to post a message, even a very short one, he/she has clearly paid some attention to the related topic. In addition, by restricting the message count to the overnight period, this attention measure corresponds directly to the overnight textual sentiment $S_{o,t}$ (recall the timeline depicted in Figure IV).

 Armed with these empirical proxies, we estimate the state-dependent effect of textual sentiment on future stock prices using the following augmented specification (cf. equation (4)) with interaction terms:

$$R_{o,t+1} = \alpha + \beta \times S_{o,t} + (\gamma_1 A_{o,t} + \gamma_2 RV_t + \gamma_3 Vlm_t + \gamma_4 LowGrowth_t) \times S_{o,t} + \delta \times Controls_t + \epsilon_{t+1},$$

where $Controls_t$ contain the full set of control variables used in columns (3) and (6) of Table V. For brevity, we only report results for the value-weighted overnight CISI sentiment index. Note that all explanatory variables on the right-hand side of equation (6) are observed before the 9:15 am call-auction session, exactly like the predictive regression in equation (4) studied above.

Table IX presents estimation results based on equation (6) and its submodels. For ease of comparison, the first column reports the benchmark specification without including any interaction terms (i.e., $\gamma = 0$), which replicates column (3) of Table V. Turning to the state-dependent estimates, we observe from column (2) of Table IX that the coefficient of the interaction term $S_{o,t} \times A_{o,t}$ is positive and statistically highly significant. This effect is also economically large: A one-standard-deviation increase in $A_{o,t}$ is associated with a 8.5-basis-point increase in the effect of sentiment on stock returns, which is comparable to the average marginal effect of 12.7 basis points. This result supports the hypothesis that stock returns are more responsive to sentiment when investors pay more attention. This finding is qualitatively unaltered when we further include the interaction term $S_{o,t} \times RV_t$ (see column (3)), the coefficient of which is also significantly positive.

24 The Google-trend-based measure is not informative for our study of China’s stock market, because Google does not provide service in mainland China. Similar services in China are much less accessible, and their quality remains to be investigated in future work.
This finding is in line with Shleifer and Vishny’s (1997) argument that risk-averse arbitrageurs may avoid volatile markets, leaving the stock price more sensitive to sentiment-driven noise trading.

By contrast, as shown by the estimates in columns (4) and (5) of Table IX, the trading volume ($Vlm_t$) and the indicator for economic downturn ($LowGrowth_t$) do not have a significant impact on the predictive return-sentiment relationship. Including the interaction term $S_{o,t} \times Vlm_t$ actually renders the coefficient of $S_{o,t}$ very noisy, which is mainly due to the high correlation between $A_{o,t}$ and $Vlm_t$ (correlation = 0.66). Meanwhile, the lack of significance of the low-growth indicator is hardly surprising, because identifying the time-varying effect at business-cycle frequency invariably demands a very long sample, such as the 100-year sample studied by García (2013). Unfortunately, such a long sample is rarely available in less-developed financial markets, such as the one in China.

In view of the above discussion, we consider column (3) of Table IX as the preferred specification. This specification implies a time-varying coefficient of textual sentiment (the subscript $I$ stands for “implied”):

$$
\beta_{I,t} = \beta + \gamma_1 \times A_{o,t} + \gamma_2 \times RV_t.
$$

Using the estimates in Table IX, we can generate estimates of the state-dependent sentiment coefficient as $\hat{\beta}_{I,t} = \hat{\beta} + \hat{\gamma}_1 \times A_{o,t} + \hat{\gamma}_2 \times RV_t$.

The time variation in $\hat{\beta}_{I,t}$ is solely driven by the attention and volatility time series in a parametric fashion. In order to examine whether this parameterization captures the actual time-varying sentiment effect sufficiently well, we also compute 250-day rolling-window estimates from the model

$$
R_{o,t+1} = \alpha_t + \hat{\beta}_{I,t} \times S_{o,t} + \hat{\delta}_t \times Controls_t + \epsilon_{t+1},
$$

for which the estimation is based on observations dated from $t - 249$ to $t$. This rolling-window estimation can be formally given a nonparametric interpretation. Econometrically speaking, the estimated coefficient $\hat{\beta}_{I,t}^\text{Roll}$ should be considered as the “local” average of the latent time-varying coefficient within the $[t - 249, t]$ estimation window. Therefore, it is not directly comparable to $\hat{\beta}_{I,t}$, but rather should be compared to the moving average of the latter, defined as

$$
\hat{\beta}_{I,t}^\text{Roll} = \hat{\beta} + \hat{\gamma}_1 \times \frac{1}{250} \sum_{j=0}^{249} A_{o,t-j} + \hat{\gamma}_2 \times \frac{1}{250} \sum_{j=0}^{249} RV_{t-j}.
$$

---

25The lack of significance of the $LowGrowth_t$ variable is not due to its binary construction. In results not reported here, we also used the level of real GDP growth and found very similar results.

26As a result of the high correlation between $A_{o,t}$ and $Vlm_t$, the regression on $S_{o,t}$, $S_{o,t} \times A_{o,t}$ and $S_{o,t} \times Vlm_t$ suffers from a severe multicollinearity problem.
In Figure V, we make this comparison by plotting the paths of \( \hat{\beta}_{Roll} \) (solid line) and its nonparametric counterpart \( \hat{\beta}_{Roll} \) (dashed line). The 95%-level confidence band of the latter is also plotted to help in assessing the sampling variability of this less precise nonparametric estimate.

Several interesting findings emerge from this figure. First, from the nonparametric rolling-window estimate, we observe some salient time variation in the responsiveness of stock returns to textual sentiment. The sentiment coefficient is high at the beginning of our sample, which corresponds to the U.S. subprime mortgage crisis, and peaks during China’s 2015–2016 stock market crash. Second, we see that the state-dependent coefficient implied by the parametric model tracks the nonparametric estimate remarkably well. The path of \( \hat{\beta}_{Roll} \) falls in the confidence band around the nonparametric estimate for a large portion of the sample, suggesting that the overnight attention measure \( A_{o,t} \) and the (log) realized volatility \( RV_t \) are sufficiently adequate for capturing the time-varying effect of textual sentiment on stock returns.

To sum up, we find that investor attention (measured by the log volume of text messages) and realized volatility are positively associated with the responsiveness of future stock price to textual sentiment. Moreover, we show that these state variables are adequate for tracing the time-varying return-sentiment relationship nonparametrically recovered from a standard rolling-window estimation. These findings are consistent with the implications of the limited-attention theory (Kahneman (1973)) and the theory of limits-to-arbitrage (Shleifer and Vishny (1997)), and suggest that these theories capture the main drivers of the time-varying effect. On the other hand, state variables like trading volume (as a conventional attention proxy) and below-median real GDP growth (as a proxy for economic downturn) do not appear to have the same empirical relevance in our analysis.

IV.E. Textual Sentiment and Volatility

An important prediction of De Long et al.’s (1990) theory is that noise trading induces excess volatility (Shiller (1981)): If noise traders base their trading decisions on sentiment, unusually high or low sentiment should induce high volatility. The financial econometrics literature on ARCH-type models (Engle (1982), Bollerslev (1986)) further argues for the empirical relevance of “asymmetric” or “leverage” volatility models, in which bad news has a larger impact on future volatility than good news; see, for example, Black (1976), Nelson (1991), Engle and Ng (1993), and Glosten et al. (1993), among others. Guided by these studies, we test two hypotheses in this subsection: the
excess-volatility hypothesis that unusually high or low sentiment predicts higher volatility, and the asymmetry hypothesis that volatility is more responsive to pessimistic sentiment than to optimistic sentiment.

The existing text-based empirical evidence regarding these hypotheses is somewhat limited and mixed. Antweiler and Frank (2004) show that their message-board bullishness measure positively predicts volatility, but find little support for the asymmetry hypothesis. Da et al. (2015) construct a sentiment measure using the Google search volume of phrases related to household concerns (e.g., “recession” and “unemployment”). The authors show that negative sentiment is associated with high volatility contemporaneously, but without significant predictive power. The most supportive evidence for both of these hypotheses is provided by Loughran and McDonald (2011), who document that both negative and positive word counts in corporate 10-K reports positively predict future volatility, with negative words exerting a stronger influence than positive ones. Loughran and McDonald’s findings are based on (low-frequency) annual observations of 10-K reports, for which the identification mainly arises from the cross-section; in contrast, we focus on the volatility of the market portfolio using (daily) high-frequency time-series data.

Our empirical tests rely on the following regression to predict the day \( t + 1 \) log realized volatility \( (RV_{t+1}) \):

\[
RV_{t+1} = \alpha + \beta_1 \times |S_{o,t}| + \beta_2 \times S_{o,t} + \beta_3 \times A_{o,t} \\
+ \gamma_1 \times |S_{o,t}| \times A_{o,t} + \gamma_2 \times S_{o,t} \times A_{o,t} + \delta \times Controls_{t} + \epsilon_{t+1}.
\]

We remind the reader that realized volatility is computed as the sum of squared 5-minute returns within the regular trading hours between 9:30 am and 3:00 pm. A few remarks on the above specification are in order. We include both \( |S_{o,t}| \) and \( S_{o,t} \) to test the excess-volatility and asymmetry hypotheses discussed above, which suggest \( \beta_1 > 0 \) and \( \beta_2 < 0 \), respectively. In view of Antweiler and Frank’s (2004) finding that the number of messages positively predicts volatility, we also include \( A_{o,t} \) as an explanatory variable. The interaction terms \( |S_{o,t}| \times A_{o,t} \) and \( S_{o,t} \times A_{o,t} \) are also included to capture the idea that the effect of sentiment on volatility may be amplified by high investor attention, corresponding to \( \gamma_1 > 0 \) and \( \gamma_2 < 0 \).

The control variable vector \( (Controls_t) \) is set as follows. As in the return regressions, we include five lagged daily returns of the CSI 300 index and the latest close-to-close returns of the S&P 500 index and NYMEX oil futures. However, due to the well-known long-memory feature of volatility, we control for volatility dynamics differently from the return regressions, by including lagged log realized volatilities up to 1 trading month (i.e., 22 days). For parsimony, we follow the HAR model (Corsi (2009)) by including daily, weekly, and monthly lags of the log realized volatility (i.e., \( RV_t \),
(1/5) ∑_{j=0}^{21} RV_{t-j} and (1/22) ∑_{j=0}^{21} RV_{t-j} in the control variables. This HAR specification provides a simple way to capture long-memory volatility dynamics, and has arguably become the most popular model in the recent financial econometrics literature for volatility forecasting. Finally, in view of the extensively documented positive relationship between volume and volatility (Clark (1973), Tauchen and Pitts (1983)), we also control for daily, weekly and monthly lags of log trading volume, mirroring the HAR specification of lagged volatilities.

Table X reports the estimation results. From columns (1)–(4), we see that the coefficients of |S_{o,t}|, S_{o,t}, and A_{o,t} are all statistically significant at the 1% level, regardless of whether we include them individually or jointly in the regressions. Importantly, the estimates strongly support the excess-volatility and asymmetry hypotheses discussed above. Focusing on the results in column (4), we see that a one-standard-deviation increase in the magnitude of overnight sentiment predicts a 7.5% increase in the next day’s log realized volatility. Further, the negative coefficient of S_{o,t} clearly indicates that the news impact curve (Engle and Ng (1993)) is asymmetric. Like Antweiler and Frank (2004), we also find that the number of messages strongly predicts high levels of volatility, with an estimated elasticity of 11.3%. Finally, as shown in column (5), these findings do not change when we further include interaction terms between sentiment and attention. We find that the effect of |S_{o,t}| on volatility is larger when investors pay more attention, but do not find the same significant evidence for the asymmetric response.

In summary, our findings from the predictive volatility regression strongly support both the excess-volatility hypothesis and the asymmetry hypothesis. To the best of our knowledge, our results are the first to support both hypotheses for the market portfolio using textual sentiment measures (cf. Antweiler and Frank (2004) and Da et al. (2015)). Our findings also complement the cross-sectional evidence provided by Loughran and McDonald (2011) for the U.S. market.

V. Sentiment, disagreement, and trading volume

In this section, we study whether extreme textual sentiment and high textual disagreement predict more trading volume. Section V.A develops related hypotheses and describes the construction of our textual disagreement measure. Section V.B presents the empirical results.

\footnote{See, for example, Andersen et al. (2007), Andersen et al. (2011), Patton and Sheppard (2015), Bollerslev et al. (2018), Bollerslev et al. (2019), and references therein.}
V.A. Hypothesis Development

Investor sentiment manifests in trading activity. The classical ideas about bubbles of Adam Smith and his contemporaries rely on the concept of “overtrading” (Kindleberger and Aliber (2005)); Baker and Stein (2004) and Baker and Wurgler (2006) propose volume as a proxy for investor sentiment.\(^{28}\) Tetlock (2007) finds that unusually high or low sentiment extracted from the Wall Street Journal predicts high volume (also see Loughran and McDonald (2011) and García (2013)). On the other hand, Hong and Stein (2007) emphasize that although sentiment-driven noise trading in classical models can generate trading volume, the amount is not always large. The authors suggest that the class of “disagreement” models can provide a robust explanation for high trading volume (Harris and Raviv (1993), Kandel and Pearson (1995)). In these models, investors interpret the same news differently, and they agree to disagree. As such, investor disagreement provides an extra trading motive.

Inspired by the prior theoretical studies, we examine two hypotheses regarding trading volume in the context of China’s stock market. One is the sentiment-volume hypothesis, by which unusually high or low sentiment predicts trading volume. The other is the disagreement-volume hypothesis, by which higher investor disagreement predicts larger trading volume. The first hypothesis can be readily investigated using the textual sentiment measure developed and studied above. To test the second hypothesis, we further construct a textual measure for investor disagreement. The idea is simple: While the textual sentiment index measures the average textual tones of different messages, we measure disagreement via their dispersion. As in our return regressions, we focus on whether overnight disagreement predicts the next day’s trading volume, following the scheme depicted in Figure IV.

We construct the overnight disagreement measure in two steps. In the first step, we follow Antweiler and Frank (2004) and construct the stock-specific overnight disagreement index for stock \(i\) on day \(t\) as

\[
D_{o,t}^{(i)} = \sqrt{1 - \frac{(P_{o,t}^{(i)} - N_{o,t}^{(i)})^2}{P_{o,t}^{(i)} + N_{o,t}^{(i)}}},
\]

where \(P_{o,t}^{(i)}\) and \(N_{o,t}^{(i)}\) are the numbers of positive and negative messages (classified by the textual analysis methods described in Appendix II.E) during the overnight period for stock \(i\), respectively. We note that \(D_{o,t}^{(i)}\) takes values between zero and one. It equals zero (i.e., no disagreement) when

\(^{28}\) In a recent update of their sentiment index, Baker and Wurgler dropped the NYSE turnover in their construction of the sentiment index, because “turnover does not mean what it once did, given the explosion of institutional high-frequency trading and the migration of trading to a variety of venues.” Also see Jeffrey Wurgler’s research webpage.
all messages are in the same sentiment category—positive or negative—and it equals one (i.e., maximal disagreement) when $P_{o,t}^{(i)} = N_{o,t}^{(i)}.$ In the second step, we compute the market-level overnight disagreement $D_{o,t}$ as the value-weighted average of these stock-specific $D_{o,t}^{(i)}$ disagreement indexes. To simplify interpretation of the regression coefficient, we standardize the $D_{o,t}$ series to have zero mean and unit standard deviation.

It is a priori unclear whether this disagreement measure can provide support for the disagreement-volume hypothesis. Antweiler and Frank (2004) construct a similar measure using messages from Yahoo! Finance and Raging Bull message boards for the year 2000, but find that, as opposed to the aforementioned disagreement-volume hypothesis, higher disagreement Granger-causes lower trading volume. We believe that Antweiler and Frank’s finding should be interpreted cautiously, because it is based on a short (1-year) sample at the peak of the dot-com bubble. For this reason, re-examination of the disagreement-volume hypothesis seems warranted.

Another possible measure for disagreement is the Economic Policy Uncertainty (EPU) index proposed in a recent influential paper by Baker et al. (2016). The EPU index may serve as a disagreement measure because, to the extent that investors are uncertain about future policies, they may interpret public news in different ways based on their own beliefs. Bollerslev et al. (2018) show that the EPU is a relevant proxy for disagreement in their study of trading volume in the U.S. stock and bond markets. Therefore, we also include the China EPU index (henceforth CEPU) in our analysis below as an alternative to our textual disagreement measure. The CEPU is based on the South China Morning Post, Hong Kong’s leading English-language newspaper, and is available at monthly frequency; we use the latest update for each day in our predictive analysis to avoid look-ahead bias.

V.B. Empirical Results on Trading Volume

We test the sentiment-volume hypothesis and the disagreement-volume hypothesis using the following predictive regression for the day $t + 1$ log trading volume ($Vlm_{t+1}$):

\begin{equation}
Vlm_{t+1} = \alpha + \beta_1 \times |S_{o,t}| + \beta_2 \times S_{o,t} + \beta_3 \times D_{o,t} + \beta_4 \times CEPU_t + \beta_5 \times A_{o,t} + \gamma_1 |S_{o,t}| + \gamma_2 S_{o,t} + \gamma_3 D_{o,t} \times A_{o,t} + \delta \times Controls_t + \epsilon_{t+1}.
\end{equation}

We set the disagreement measure to zero if the numbers of positive and negative messages are both zero. The $D_{o,t}^{(i)}$ index is a dispersion measure in the following mathematical sense. Consider a random variable $X$ that equals to 1 with probability $P_{o,t}^{(i)}/(P_{o,t}^{(i)} + N_{o,t}^{(i)})$ and -1 otherwise. The measure $D_{o,t}^{(i)}$ is the standard deviation of $X$.

Yet another type of proxies for disagreement is the dispersions in the survey of professional forecasters (SPF) for various macroeconomic variables (see, e.g., Ilut and Schneider (2014)). Unfortunately, a long and reliable record of surveys like the SPF is unavailable for China, which is likely the case for other less developed economies as well.

Data are obtained from the Economic Policy Uncertainty website, labeled the “China Monthly Index.” Source: http://www.policyuncertainty.com/china_monthly.html
Under the sentiment-volume hypothesis, the coefficient of $|S_{o,t}|$ should be positive (i.e., $\beta_1 > 0$). The $S_{o,t}$ variable is also included to potentially capture any asymmetric response of volume with respect to textual sentiment. Meanwhile, if the disagreement-volume hypothesis is in force, we expect the $\beta_3$ and $\beta_4$ coefficients to be positive, provided that the textual disagreement $D_{o,t}$ and the CEPU are indeed relevant proxies of investor disagreement. Since the effects of these variables on the next day’s volume may be amplified when investors pay more attention, we include the interaction terms of these variables with the attention measure $A_{o,t}$, and the two hypotheses imply $\gamma_1 > 0$ and $\gamma_3 > 0$. Antweiler and Frank (2004) find that message posting activity positively predicts volume; hence, we also include the $A_{o,t}$ variable by itself and expect the coefficient $\beta_5$ to be positive, which is also in line with the early theoretical models of Clark (1973), Tauchen and Pitts (1983), and Andersen (1996). Finally, to account for the high persistence of the volume series, we use the exact same set of control variables as in the volatility regression specified in equation (9), which, in particular, includes both lagged volume and realized volatility up to 1 month using the HAR specification.

[INSERT TABLE XI]

Table XI presents the estimation results. Columns (1)–(5) report the marginal effects of individual variables. We mainly discuss the estimation results for the full specification in column (6) for brevity. First, we find strong support for the sentiment-volume hypothesis, as the $\beta_1$ and $\gamma_1$ coefficients are both positive and highly statistically significant. These effects are also economically large: On average, a one-standard-deviation increase in the magnitude of sentiment predicts a 2.5-percentage-point increase in log volume, and this effect is more than doubled when the attention measure is one standard deviation above average. We find little evidence for asymmetric response, as the $\beta_2$ and $\gamma_2$ coefficients are both close to zero, which is unlike the pronounced asymmetry pattern seen in the volatility regressions above. This contrast is interesting, because it suggests that the well-known volume-volatility relationship may depend on the state of investor sentiment.\footnote{Related to this finding, Bollerslev et al. (2018) document that the volume-volatility elasticity in the U.S. stock and treasury bond markets is lower (i.e., closer to zero) when the textual tone of the actual FOMC statement is more negative.}

Overall, these supportive findings for the sentiment-volume hypothesis confirm and extend prior empirical results for U.S. markets (Tetlock (2007), Loughran and McDonald (2011), García (2013)). In addition, the positive $\beta_5$ coefficient suggests that high activity on the message board predicts high volume, which is consistent with Antweiler and Frank (2004).

Our more unique finding in Table XI pertains to the effect of investor disagreement. From column (3), we observe that the coefficient of the textual disagreement $D_{o,t}$ is positive and marginally
statistically significant. This finding provides some moderate support for the disagreement-volume hypothesis. In contrast, we see from column (4) that the coefficient of CEPU is very close to zero and statistically insignificant. These findings still hold when we control for other variables in column (6). Interestingly, under this full specification, we see that the positive effect of the textual disagreement \( D_{o,t} \) mainly stems from its interaction with the attention measure \( A_{o,t} \), as the coefficient of their interaction term is highly significant. This finding is intuitive, in that the disagreement can only exert influence on trading when investors pay enough attention to the underlying news that induced their disagreement in the first place. Hence, it is reasonable to see that the disagreement-volume hypothesis manifests more when investor attention is higher. However, the coefficient of CEPU is, again, close to zero. This shows that the CEPU does not capture investor disagreement well in the current context.\(^{33}\)

All in all, our empirical findings provide strong support for the sentiment-volume hypothesis, by which unusually high or low textual sentiment predicts high volume. Consistent with the key implications of the disagreement models (see Hong and Stein (2007) and references therein), we also show that higher textual disagreement predicts higher trading volume; to the best of our knowledge, this is the first result of its kind in the literature, and it differs from Antweiler and Frank’s (2004) earlier result that higher textual disagreement predicts lower trading volume. Future work on different stock markets (including the U.S. market over a longer time period) is needed to further test the implications of disagreement models.

### VI. Conclusion

We construct textual sentiment measures for China’s stock market using a unique—and massive—online message dataset for the period between 2008 and 2018. Our construction relies on both conventional dictionary-based methods and state-of-the-art machine-learning tools from computational linguistics. Combining these methods, we build the China Investor Sentiment Index (CISI) as a barometer of Chinese investors’ sentiment, and quantify its effect on future market returns, volatility, and trading volume. First and foremost, we document that textual sentiment can significantly predict market returns, exhibiting a salient underreaction-overreaction pattern on a time scale of several months. The impact of textual sentiment is larger and longer lasting for small (resp. growth) stocks than big (resp. value) stocks. We also document the time variation in the return-sentiment relationship by showing that textual sentiment matters more when investor attention is

---

\(^{33}\)There are several reasons why the CEPU may be less informative than the U.S. EPU index: The CEPU is based on only one newspaper, it is not written in the native language, and it may be subject to regulations that differ from those in the U.S.
higher and when markets are more volatile. This evidence supports theoretical predictions regarding noise-trading, limits-to-arbitrage, and limited attention, as emphasized in behavioral economics. In addition, we find that unusually high and low textual sentiment predicts higher volatility, but negative sentiment has a much larger effect. These findings are consistent with predictions that sentiment can generate excess volatility, and its impact is asymmetric. Lastly, we also find that higher trading volume is predicted by extreme sentiment and more differences of opinion as measured by textual disagreement. This finding supports the classical idea of “overtrading” and the central prediction of disagreement models, in which heterogeneous agents agree to disagree.

The paper’s contribution can be summarized as follows. Our unique textual dataset allows us to test a broad range of hypotheses from behavioral finance within a unified empirical framework. As a by-product, it also allows us to access the relative merits of alternative textual analysis approaches in an economically important context, which ought to be informative for future research in the burgeoning textual analysis literature of economics and finance. Our empirical findings provide support for core ideas in behavioral finance, confirming and extending evidence in the text-based empirical literature. As a more specific contribution, this is the first comprehensive study on textual sentiment in China’s enormous stock market. Our empirical study not only provides a better understanding of the financial market of the largest developing country per se, but also provides an out-of-sample check for related empirical findings in the literature, which are mainly based on the U.S. market. We believe that our textual sentiment (and disagreement) indexes can be used to study a variety of interesting questions in future research.
Appendix

I Protocol of the labeling process

We randomly select 200 messages from each of the 200 largest (by market capitalization) stocks to form the training subsample of 40,000 messages. The team of readers consists of fourteen graduate students majored in economics at Peking University. The task is to classify each message into one of the following three categories: “positive,” “negative” or “ambiguous.” The readers also need to provide a list of keywords that inform their judgments. We provide a four-hour training session to the readers, in which we describe a detailed guideline with many examples. The guideline is summarized as follows.

Positive. A message is labeled “positive” if the investor (1) posts news announcement that delivers positive sentiment; (2) reveals his/her intention to buy the stock; (3) recommends buying this stock; (4) claims that he/she has already bought this stock; (5) expresses negative attitude toward those who sell/short this stock; (6) believes that the stock price will rise in the near future; (7) is optimistic about the company’s earnings prospect; (8) shows off his earnings from buying this stock; or (9) compliments on other aspects of this company.

Negative. A message is labeled “negative” if the investor (1) posts a news announcement that delivers negative sentiment; (2) reveals his/her intention to sell the stock; (3) advises others to sell the stock; (4) predicts that the stock price will fall in the near future; (5) is pessimistic about the company’s earnings prospect; (6) is disappointed by dividend payment; (7) curses; (8) complains about his/her losses from buying this stock, or (9) complains about other aspects of this company.

Ambiguous. A message is labeled “ambiguous” if (1) the message appears to be incomplete; (2) the message is an advertisement unrelated to the stock market; (3) the investor raises a question; (4) the investor describes his/her situation and seeks advice; (5) the investor provides a plan that is contingent on the performance of a stock or the market; (6) other ambiguous messages.

After the training session, each reader needs to pass a related test before starting the actual labeling task. The labeling process is then carried out in two stages. In the first stage, we assigned 1,000 messages to each reader. The readers were informed that their output would be audited randomly and their payments would depend on the quality of their output. After the auditing, we retain the better-performing readers who then labeled the remaining messages in the training sample. As mentioned in the main text, each message was independently read by two readers. We deem the labels valid only when two readers provide the same labels (recall footnote 13 for our rationale). On average, a reader spent 51 seconds on each message; the total time cost of
this labeling process is approximately 1,200 hours. We randomly audited 2,500 messages in total and manually screened all the keywords identified by the readers. The final training set contains 27,999 valid labeled messages. The keywords are used to build the Chinese Stock Market (CSM) dictionary.

II Technical details on textual analysis

In Appendices II.A–II.D, we describe the computational details for (1) parsing the text messages; (2) the vector representation of words using word2vec; (3) implementation of the support vector machine; and (4) implementation of the convolutional neural network. In Appendix II.E, we present results on the comparison among different textual analysis methods.

II.A. Text Segmentation

Unlike Germanic and Latin languages, Chinese words in a sentence are not separated by spaces. We use the Python package jieba to parse text messages into individual words. This package is often used for parsing Chinese texts in the computational linguistics community. The jieba package performs the parsing task based on a dictionary that specifies the different words. Jieba’s default dictionary contains most of the commonly used Chinese words, but a user can provide supplemental dictionaries tailored for specific applications. We augment jieba’s default dictionary with three supplements: (1) our CSM dictionary; (2) a dictionary containing the names of listed firms; and (3) a generic dictionary of financial terms from the Sougou lexicon.

As English, the Chinese language also contains “stop” words. These very common words are typically discarded in textual analysis, which can be done directly using jieba in the parsing step. We remove 17 stop words that are the Chinese equivalents of the following words: of, -ing, -ed, and, is/are, about, both/all, yet, or, a/an, we/us, you, they/them. (There are three versions of “and” and three versions of “they/them.”) In addition, we also remove the very rare words that appear less than 10 times in our entire sample (which consists of approximately 60 million messages). This parsing process results in a vocabulary of 654,555 words in total.

---

34 The jieba package is available at https://github.com/fxsjy/jieba.
35 The lexicon is available at https://pinyin.sogou.com/dict/detail/index/15127?rf=dictindex. Sougou is the leading company for developing Chinese input softwares.
II.B. Vector Representation of Words

In the context of this paper, a machine-learning method (SVM or CNN) takes each text message as the input and generates the associated category-specific probabilities (and thus the sentiment score defined in equation (3)) as the output. An important component in the implementation is to represent words as numeric vectors. Two representation methods are commonly used: one-hot and word2vec. The former is a simple method based on word counts, whereas the latter is a state-of-the-art method developed in the recent computational linguistics literature. Although we only use the latter in the construction of the textual sentiment index, we also experimented with the one-hot representation when implementing the SVM and found that it underperforms the word2vec representation (see Appendix II.E below). In this subsection, we describe the one-hot and the word2vec representations in detail.

We start with the one-hot representation. Let $n$ be the size of the vocabulary, that is, the number of distinct words used by all messages in the dataset (recall that $n = 654,555$ in our sample). Under the one-hot representation, each text message is represented as a $n$-dimensional vector, whose $k$th element equals the times that the $k$th word in the vocabulary appears in this message. Since most words in the vocabulary do not appear in each individual message, the corresponding vector is high-dimensional, but with most elements being zero. As a concrete illustration of the one-hot representation, consider a sample with two messages: “apple, orange, apple” and “apple, cake.” The vocabulary is \{apple, cake, orange\}. The one-hot representation of the two messages are (2,0,1) and (1,1,0), respectively.

The word2vec method transforms each word into a $K$-dimensional real vector (with $K \ll n$), so that semantically close words correspond to numerically close vectors. Following Mikolov et al. (2013) and Mikolov et al. (2013), we implement the word2vec representation using the skip-gram model. The idea underlying this model is to use each word to predict its surrounding words, and vector representations are chosen to maximize the likelihood function of this prediction problem.

We implement the word2vec representation using the function `Word2Vec` provided in the Python package `gensim` by Řehůrek and Sojka (2010).\(^{37}\) The exact implementation is carried out as follows. The input to the `Word2Vec` function is our parsed message data with stop words, punctuation, numbers, and very rare words removed, as described in the previous subsection. We set the window width to 10, which specifies the maximum distance between each word and its surrounding words to be predicted in the skip-gram model. The other tuning parameters are fixed at their built-in default values.

II.C. Support Vector Machines

In this subsection, we describe how to construct sentiment scores of individual text messages using support vector machines (SVM). We first briefly discuss the SVM algorithm and then describe how we implement the computation using the sklearn package in Python. We refer the reader to Chapter 9 of James et al. (2013) and Chapter 12 of Trevor et al. (2017) for additional details on SVMs and references therein.

We represent each message $i$ in the sample as $(x_i, y_i)$: The “label” variable $y_i$ takes value in \{positive, negative, ambiguous\} and the “feature” variable $x_i$ is a vector representation of the text message. In our application, $x_i$ is the average of the word vectors in this message, with the latter computed using either the one-hot or the word2vec (with $K = 200$) representation. Note that $y_i$ is observed only if message $i$ has been labeled by a human reader. SVMs serve the purpose of estimating the conditional probabilities that a generic (labeled or unlabeled) message falls in one of the three categories given its observed feature $x_i$, that is,

$$P(y_i = c|x_i), \quad c \in \{\text{positive, negative, ambiguous}\}. \quad (12)$$

The sentiment score is then given by

$$P(y_i = \text{positive}|x_i) - P(y_i = \text{negative}|x_i).$$

Before diving into the computational details, it is instructive to briefly review the working of SVMs. In its basic form, a SVM concerns a two-class classification problem in which $y_i$ takes value in, say \{-1, 1\}, without loss of generality. The classifier uses a simple decision rule: $y_i = 1$ (resp. $-1$) if $\beta_0 + x_i^T \beta$ is positive (resp. negative); the parameters $\beta_0$ and $\beta$ are chosen so that “most” observations in the labeled sample $(x_i, y_i)_{1 \leq i \leq N}$ are correctly classified. More precisely, $\beta_0$ and $\beta$ are obtained by solving the following minimization (see equation (12.8) in Trevor et al. (2017)):

$$\min_{\beta_0, \beta, \xi_1, \ldots, \xi_N} \frac{1}{2} \|eta\|^2 + C \sum_{i=1}^{N} \xi_i$$

subject to $$\begin{cases} 
\xi_i \geq 0, \\
y_i (\beta_0 + x_i^T \beta) \geq 1 - \xi_i, 
\end{cases}$$

for all $1 \leq i \leq N$,

where $\|\beta\|$ denotes the Euclidean norm of $\beta$, $C$ is the user-specified “cost” parameter, and the $\xi_i$’s are slack variables that are determined by the minimization. Intuitively, the value of $\xi_i$ quantifies the extent to which the $i$th observation is misclassified, and the misclassification is more heavily
penalized when the $C$ parameter is larger. As such, the choice of $C$ reflects the usual bias-variance trade-off, with a large or small value resulting in over- or under-fitting, respectively. In practice, this tuning parameter is routinely chosen via cross-validation.

Compared with this “textbook” version of SVM, our application involves two additional complications. First, we consider a three-class problem rather than the basic two-class problem. A standard way of addressing this issue (see Section 9.4 of James et al. (2013)) is the “one-versus-one” procedure, in which we first conduct two-class classifications for all label pairs (i.e., \{positive, negative\}, \{positive, ambiguous\}, \{negative, ambiguous\}) and then aggregate these pairwise classification results. Second, the direct output of SVM is the predicted label, rather than the category-specific conditional probabilities. But this issue can also be addressed using standard methods. In the two-class problem, the probabilities can be computed using Platt scaling (Platt (1999)). Here, we adopt the multi-class extension given by Wu et al. (2004).

Our numerical implementation is carried out straightforwardly using the $SVC$ function in the Python package $sklearn$.\textsuperscript{38} More specifically, we use a linear kernel by setting the “kernel” option to “linear.” We set the “multi_class” option to “ovo” to implement the one-versus-one multi-class classification. Finally, we set the “probability” option to “True” for obtaining the category-specific conditional probabilities. The cost parameter is set to $C = 2$ as determined by cross-validation.

\textit{II.D. Convolutional Neural Networks}

In this subsection, we provide a brief discussion on convolutional neural networks (CNN) and describe how the CNN method is implemented in our analysis. CNNs have been successfully applied in computer vision (Krizhevsky et al. (2012)) and speech recognition (Graves et al. (2013)), and remain at the forefront of commercial applications of deep-learning methods (Goodfellow et al. (2016)). Kim (2014) introduces CNN to textual analysis. In our application, we adopt a similar strategy for computing the conditional probabilities in equation (12), from which we compute the CNN-based sentiment scores.

CNN can be considered as a type of Artificial Neural Networks (ANN), which have been studied by econometricians since Gallant and White (1988) and Hornik et al. (1989); also see Chen (2007) for a more recent review. For classification problems, ANNs are akin to nonlinear logit models. Compared with standard ANNs, a CNN involves a preliminary convolution step, as its names suggests. In the textual analysis context, the convolution amounts to applying “filters” on the textual data that compute locally weighted averages of the (vector representations of) adjacent

\textsuperscript{38}Source: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC.
words in the text message. By using different local window sizes and weighting functions, the convolution can help uncover useful textual features at different “resolutions.” These extracted features, rather than the words themselves, are then used as inputs of the neural network. (This is analogous to looking at “patterns” in a picture rather than zooming into the individual pixels.)

Of course, the practical implementation of CNN can vary significantly across different applications. We refer the reader to Goodfellow et al. (2016) for a more comprehensive discussion on CNNs. The discussion here is to clarify our exact implementation so as to facilitate replication and extension in future work. We conduct the computation using the Python deep-learning library *keras*.39 Below, we itemize the main components of the algorithm with associated implementation details. Whenever possible, we intentionally replace machine-learning jargons with their econometric analogues.

*Representation of the textual message.* We parse each text message as described in Appendix II.A and represents each word as a $K = 50$ dimensional vector obtained using word2vec (recall Appendix II.B). Each text message can thus be viewed as an (ordered) “time series” $x_1, \ldots, x_m$ of $K$-dimensional vectors, where $x_i$ corresponds to the $i$th word in the parsed text. For computational tractability, we only keep the first 100 words of each message.

*Convolution.* The convolution step involves computing the weighted moving averages of the word vector series with different window sizes and weights. For example, when the window size is 3 and the weights are $(w_1, w_2, w_3)$, the output of the convolution is the moving average $\sum_{j=1}^{3} w_j x_{i+j}$ for $i \geq 1$. Following Kim (2014), we use three window sizes (3, 4, and 5) and, for each window size, we consider 25 different sets of weights; this results in $3 \times 25 = 75$ filters in total. The weight parameters (i.e., the $w$’s) are estimated in the training step through optimization. The number of weighting schemes (i.e., 25) is a tuning parameter determined adaptively by cross-validation. These calculations are implemented using the *Conv1D* function in the *keras* package; we set the “activation” option to “relu” for using the rectified linear unit (ReLU) activation, which is a standard choice.

*Pooling.* Almost all CNNs involve a pooling step for further modifying the output from the convolution step. This makes the extracted features less sensitive to the exact positions of the words. We adopt a local max pooling scheme, which transforms the moving average series (from the convolution step) into their local maximum (with the same window size). The pooling step is implemented using the *MaxPooling1D* function.

*Classification via neural network.* The features computed from the pooling step are then used

---

39Source: https://keras.io/.
as the input of a neural network. We use a two-layer neural network. The first layer contains 50 neurons with the ReLU activation. The second layer is fully connected with the softmax activation function for the three-class (i.e., positive, negative, ambiguous) classification. In addition, we apply several standard techniques for reducing overfitting and improving computational efficiency: (1) we randomly drop half of input variables by setting the dropout rate to 50%; (2) we set the batch size to 128 and train the CNN with 20 epochs allowing for early stopping. The optimization is done using the categorical cross-entropy loss function and the RMSProp algorithm.

II.E. Comparison of Textual Analysis Methods

In this subsection, we compare the classification accuracy of different textual analysis methods. We consider seven methods. Four of them have been described in Section III.A of the main text, which are based on the Chinese LM dictionary, the CSM dictionary, the word2vec-based SVM, and the word2vec-based CNN. We note that the dictionary methods are implemented using the so-called TFIDF (term frequency inverse document frequency) weighting through the transformation in equation (1). We thus denote the two dictionary methods as LM–TFIDF and CSM–TFIDF, respectively.

We also consider three additional methods that have been used in the literature. One is the SVM method based on the one-hot representation (recall Appendix II.B). The other two methods are based on the two dictionaries under equal weighting (EW), instead of the TFIDF weighting, which compute message $i$’s sentiment score using the word counts $W_{i,j}$ directly (cf. equation (2)):

$$\text{score}_i = \frac{\sum_{j \in \text{PosList}} W_{i,j} - \sum_{j \in \text{NegList}} W_{i,j}}{\sum_{j \in \text{PosList}} W_{i,j} + \sum_{j \in \text{NegList}} W_{i,j}}.$$

We denote these methods as LM–EW and CSM–EW, respectively.

Under the dictionary methods, we classify a message according to the following decision rule

$$\text{classify message } i \text{ as } \begin{cases} \text{positive} & \text{if } \text{score}_i > 0, \\ \text{negative} & \text{if } \text{score}_i < 0, \\ \text{ambiguous} & \text{if } \text{score}_i = 0. \end{cases}$$

Note that a message that contains neither positive nor negative word will have a zero sentiment score and will be classified as ambiguous. Under the machine-learning methods, we classify message $i$ into class $c \in \{\text{positive, negative, ambiguous}\}$ if the conditional probability $P(y_i = c|x_i)$ is the highest among all three categories (recall equation (12)).
We carry out the comparison in a routine out-of-sample evaluation framework. Following standard practice in the machine-learning literature, we randomly divide the labeled sample into three subsamples for training (60%), validation (20%) and testing (20%). The training sample is used to estimate the parameters in SVM and CNN. The validation sample is used to pin down the tuning parameters, including the cost parameter $C$ in SVM, and the number of filters and the early stopping point in CNN. The remaining testing sample is used to compute the classification accuracy.

Table A.1 presents the classification accuracies for the seven aforementioned textual analysis methods. In addition to the overall accuracy (the “All” column) for all three types (i.e., positive, negative, ambiguous), we also separately report the classification accuracies for positive and negative messages in the columns labeled “Positive” and “Negative,” respectively.

We summarize the results as follows. First, among the four dictionary-based methods, CSM–TFIDF performs the best. Its overall classification accuracy is 76%, which is higher than the 71% accuracy of its equally-weighted counterpart CSM–EW. The methods based on the Chinese LM dictionary, LM–EW and LM–TFIDF, have accuracies only around 33%. These findings justify using CSM–TFIDF as the “representative” dictionary method, as described in the main text.

Turning to the performance of the machine-learning methods, we find that the one-hot-based SVM performs only slightly better than the dictionary-based CSM–TFIDF, with a two-percentage-point improvement on the classification accuracy. The main improvement arises from using the word2vec representation. As mentioned in Section III.A of the main text, the word2vec-based SVM attains a 81% classification accuracy and, incidentally, the accuracy of the word2vec-based CNN is also 81%. These findings confirm the merit of the word2vec representation as documented in the recent literature on natural language processing in other contexts. It is also interesting to note that the accuracy of these methods can reach as high as 90% for negative messages.

As a further robustness check, we repeat the evaluation analysis under an alternative sample-splitting scheme in which the training, validation and testing subsamples are ordered in time. Table A.2 presents the results. Our main qualitative findings are unaltered. First, CSM–TFIDF is the best performing dictionary method. Second, the word2vec-based SVM is similar to the word2vec-based CNN, and outperforms the one-hot-based SVM. The similarity between the findings in Table
A.1 and Table A.2 suggests that the textual structure underlying our 10-year sample is relatively stable.
References


<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>S.D.</th>
<th>Min</th>
<th>1%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>99%</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Messages (daily)</td>
<td>17,029</td>
<td>13,728</td>
<td>770</td>
<td>1,468</td>
<td>5,764</td>
<td>15,072</td>
<td>22,634</td>
<td>74,409</td>
<td>99,629</td>
</tr>
<tr>
<td>Message Length (Character)</td>
<td>110</td>
<td>428</td>
<td>1</td>
<td>3</td>
<td>15</td>
<td>26</td>
<td>48</td>
<td>703</td>
<td>421,073</td>
</tr>
</tbody>
</table>

*Notes:* This table reports summary statistics for the number of message observations per day and the message length. We report the sample mean, standard deviation (S.D.), and quantiles of each variable. The sample contains messages for CSI 300 constituent stocks during the sample period from July 1, 2008 to February 14, 2018.
## TABLE II
**Correlation Matrix of Daily Sentiment Indexes**

<table>
<thead>
<tr>
<th></th>
<th>Value-weighted</th>
<th></th>
<th>Equally weighted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dictionary</td>
<td>SVM</td>
<td>CNN</td>
<td>CISI</td>
</tr>
<tr>
<td>Dictionary</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>0.895</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNN</td>
<td>0.874</td>
<td>0.963</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>CISI</td>
<td>0.952</td>
<td>0.982</td>
<td>0.975</td>
<td>1.000</td>
</tr>
</tbody>
</table>

*Notes:* This table reports the correlation coefficients between the daily Dictionary, SVM, CNN and CISI textual sentiment indexes, where the CISI is defined as the average of the other three indexes. The left (resp. right) panel reports the correlations of the value-weighted (resp. equally weighted) versions of these indexes.
### TABLE III
**Summary Statistics of Textual Sentiment and Market Activities**

<table>
<thead>
<tr>
<th></th>
<th>CISI</th>
<th>Return</th>
<th>Volume</th>
<th>Volatility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.000</td>
<td>0.016</td>
<td>22.823</td>
<td>0.120</td>
</tr>
<tr>
<td>S.D.</td>
<td>1.000</td>
<td>1.679</td>
<td>0.646</td>
<td>1.019</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.020</td>
<td>-0.507</td>
<td>0.675</td>
<td>0.300</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>0.127</td>
<td>4.069</td>
<td>0.521</td>
<td>0.105</td>
</tr>
<tr>
<td>Corr. with CISI</td>
<td>0.240</td>
<td>-0.083</td>
<td>-0.108</td>
<td></td>
</tr>
</tbody>
</table>

**Autocorrelation Coefficients**

<table>
<thead>
<tr>
<th>Lag</th>
<th>CISI</th>
<th>Return</th>
<th>Volume</th>
<th>Volatility</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.643</td>
<td>0.034</td>
<td>0.936</td>
<td>0.820</td>
</tr>
<tr>
<td>2</td>
<td>0.552</td>
<td>-0.034</td>
<td>0.909</td>
<td>0.788</td>
</tr>
<tr>
<td>3</td>
<td>0.518</td>
<td>0.002</td>
<td>0.892</td>
<td>0.765</td>
</tr>
<tr>
<td>4</td>
<td>0.482</td>
<td>0.064</td>
<td>0.876</td>
<td>0.744</td>
</tr>
<tr>
<td>5</td>
<td>0.541</td>
<td>-0.002</td>
<td>0.862</td>
<td>0.728</td>
</tr>
<tr>
<td>22</td>
<td>0.318</td>
<td>-0.001</td>
<td>0.741</td>
<td>0.602</td>
</tr>
</tbody>
</table>

**Notes:** This table reports summary statistics of the daily textual sentiment index (value-weighted CISI), close-to-close log-return (in percentage points), log trading volume (in shares), and log realized volatility for the CSI 300 index. The sample period is from July 1, 2008 to February 14, 2018.
### TABLE IV
Baseline Predictive Return-Sentiment Relationship

<table>
<thead>
<tr>
<th></th>
<th>Dictionary</th>
<th>SVM</th>
<th>CNN</th>
<th>CISI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A1. Value-weighted Overnight Sentiment Index</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_{o,t}$</td>
<td>0.187***</td>
<td>0.186***</td>
<td>0.178***</td>
<td>0.193***</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.020)</td>
<td>(0.019)</td>
<td>(0.020)</td>
</tr>
<tr>
<td>Adj-$R^2$(%)</td>
<td>6.1</td>
<td>6.1</td>
<td>5.5</td>
<td>6.5</td>
</tr>
<tr>
<td><strong>Panel A2. Equally Weighted Overnight Sentiment Index</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_{o,t}$</td>
<td>0.201***</td>
<td>0.191***</td>
<td>0.181***</td>
<td>0.207***</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.021)</td>
<td>(0.020)</td>
<td>(0.020)</td>
</tr>
<tr>
<td>Adj-$R^2$(%)</td>
<td>7.1</td>
<td>6.4</td>
<td>5.8</td>
<td>7.5</td>
</tr>
<tr>
<td><strong>Panel B1. Value-weighted Lagged Daily Sentiment Index</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_t$</td>
<td>0.009</td>
<td>0.032*</td>
<td>0.031</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>(0.018)</td>
<td>(0.019)</td>
<td>(0.020)</td>
<td>(0.019)</td>
</tr>
<tr>
<td>Adj-$R^2$(%)</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td><strong>Panel B2. Equally Weighted Lagged Daily Sentiment Index</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_t$</td>
<td>-0.004</td>
<td>0.022</td>
<td>0.019</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.018)</td>
<td>(0.020)</td>
<td>(0.019)</td>
</tr>
<tr>
<td>Adj-$R^2$(%)</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

**Notes:** This table reports the estimation results from regressing the overnight return on the overnight or the lagged daily textual sentiment index. The overnight return $R_{o,t+1}$ is the close-to-open (3:00 pm–9:30 am) return of the CSI 300 index between day $t$ and day $t+1$. The overnight textual sentiment $S_{o,t}$ is formed using messages posted between 3:00 pm of day $t$ and 9:15 am of day $t+1$, whereas the lagged daily sentiment $S_t$ is formed using messages posted between 3:00 pm of day $t-1$ and 3:00 pm of day $t$. The four columns report the estimated sentiment coefficients using the Dictionary, SVM, CNN, and CISI textual sentiment indexes, respectively. All textual sentiment indexes are standardized to have zero mean and unit standard deviation. Standard errors displayed in parentheses are computed using the Newey–West estimator with eight lags; *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from July 1, 2008 to February 14, 2018.
TABLE V
PREDICTIVE RETURN-SENTIMENT RELATIONSHIP WITH CONTROLS

<table>
<thead>
<tr>
<th></th>
<th>Value-weighted C ISI</th>
<th></th>
<th>Equally Weighted C ISI</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_{o,t}$</td>
<td>0.187***</td>
<td>0.110***</td>
<td>0.113***</td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td>(0.020)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>$R_{t}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.007</td>
<td>0.001</td>
<td>-0.001</td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td>(0.019)</td>
<td>(0.019)</td>
</tr>
<tr>
<td>$Vlm_{t}$</td>
<td>-0.043</td>
<td>-0.023</td>
<td>-0.091</td>
</tr>
<tr>
<td></td>
<td>(0.034)</td>
<td>(0.031)</td>
<td>(0.131)</td>
</tr>
<tr>
<td>$RV_{t}$</td>
<td>-0.039**</td>
<td>-0.041***</td>
<td>-0.073</td>
</tr>
<tr>
<td></td>
<td>(0.016)</td>
<td>(0.014)</td>
<td>(0.046)</td>
</tr>
<tr>
<td>$R_{US,t}$</td>
<td>0.320***</td>
<td>0.320***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td>(0.025)</td>
<td></td>
</tr>
<tr>
<td>$R_{Oil,t}$</td>
<td>0.014*</td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.008)</td>
<td></td>
</tr>
<tr>
<td>Additional Controls</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Adj-$R^2$(%)</td>
<td>7.0</td>
<td>36.2</td>
<td>37.6</td>
</tr>
</tbody>
</table>

Notes: This table reports the results from regressing the overnight return ($R_{o,t+1}$) on the overnight CISI textual sentiment ($S_{o,t}$), value-weighted or equally weighted, while controlling for the lagged return ($R_{t}$), the log volume ($Vlm_{t}$), the log realized volatility ($RV_{t}$) for the CSI 300 index, and the latest daily returns of the S&P 500 index ($R_{US,t}$) and NYMEX oil futures ($R_{Oil,t}$). Realized volatility of the CSI 300 index is computed as the sum of squared 5-minute intraday returns. Additional control variables include lagged daily return, log trading volume, and log realized volatility variables of the CSI 300 index from day $t-4$ to day $t-1$, and weekday dummies. All textual sentiment indexes are standardized to have zero mean and unit standard deviation. Standard errors displayed in parentheses are computed using the Newey–West estimator with eight lags; *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from July 1, 2008 to February 14, 2018.
### TABLE VI
**Multi-horizon Return Regressions**

#### Panel A. Value-weighted Sentiment Index

<table>
<thead>
<tr>
<th>Return Horizon</th>
<th>1 Day</th>
<th>1 Week</th>
<th>2 Weeks</th>
<th>3 Weeks</th>
<th>4 Weeks</th>
<th>5 Weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{o,t}$</td>
<td>0.128***</td>
<td>0.273</td>
<td>0.516*</td>
<td>0.686**</td>
<td>0.883**</td>
<td>1.208**</td>
</tr>
<tr>
<td></td>
<td>(0.045)</td>
<td>(0.167)</td>
<td>(0.271)</td>
<td>(0.342)</td>
<td>(0.410)</td>
<td>(0.489)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Return Horizon</th>
<th>6 Weeks</th>
<th>10 Weeks</th>
<th>20 Weeks</th>
<th>30 Weeks</th>
<th>40 Weeks</th>
<th>50 Weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{o,t}$</td>
<td>1.515***</td>
<td>2.204***</td>
<td>2.367**</td>
<td>0.973</td>
<td>1.698</td>
<td>1.425</td>
</tr>
<tr>
<td></td>
<td>(0.558)</td>
<td>(0.681)</td>
<td>(0.943)</td>
<td>(1.144)</td>
<td>(1.295)</td>
<td>(1.308)</td>
</tr>
</tbody>
</table>

#### Panel B. Equally Weighted Sentiment Index

<table>
<thead>
<tr>
<th>Return Horizon</th>
<th>1 Day</th>
<th>1 Week</th>
<th>2 Weeks</th>
<th>3 Weeks</th>
<th>4 Weeks</th>
<th>5 Weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{o,t}$</td>
<td>0.104**</td>
<td>0.251*</td>
<td>0.478*</td>
<td>0.711**</td>
<td>0.892**</td>
<td>1.098**</td>
</tr>
<tr>
<td></td>
<td>(0.047)</td>
<td>(0.149)</td>
<td>(0.255)</td>
<td>(0.328)</td>
<td>(0.406)</td>
<td>(0.486)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Return Horizon</th>
<th>6 Weeks</th>
<th>10 Weeks</th>
<th>20 Weeks</th>
<th>30 Weeks</th>
<th>40 Weeks</th>
<th>50 Weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{o,t}$</td>
<td>1.517***</td>
<td>2.334***</td>
<td>1.768*</td>
<td>-0.032</td>
<td>0.628</td>
<td>0.666</td>
</tr>
<tr>
<td></td>
<td>(0.553)</td>
<td>(0.747)</td>
<td>(0.964)</td>
<td>(1.221)</td>
<td>(1.356)</td>
<td>(1.383)</td>
</tr>
</tbody>
</table>

**Notes:** This table reports the estimated sentiment coefficient $\beta_\tau$ in equation (5) for the specification $P_{\tau} - P_{3:00\text{pm\ day}\: t} = \alpha_\tau + \beta_\tau \times S_{o,t} + \delta_\tau \times \text{Control}_t + \varepsilon_\tau$ with $\tau$ ranges from 1-day to 1-year horizon (i.e., 50 trading weeks). The control variables include lagged daily CSI returns, log volume, and log realized volatilities up to 5 days, the lagged daily returns of the S&P 500 index and of NYMEX oil futures, and weekday dummies. Panels A and B report the results of the value-weighted and equally weighted CISI indexes, respectively. All textual sentiment indexes are standardized to have zero mean and unit standard deviation. Standard errors displayed in parentheses are computed using the Newey–West estimator with eight lags; *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from July 1, 2008 to February 14, 2018.
TABLE VII
MULTI-HORIZON REGRESSIONS FOR THE SMALL-MINUS-BIG PORTFOLIO

Panel A. Value-weighted Sentiment Index

<table>
<thead>
<tr>
<th>Return Horizon</th>
<th>1 Day</th>
<th>1 Week</th>
<th>2 Weeks</th>
<th>3 Weeks</th>
<th>4 Weeks</th>
<th>5 Weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{o,t}$</td>
<td>0.059**</td>
<td>0.224**</td>
<td>0.319</td>
<td>0.446*</td>
<td>0.396</td>
<td>0.494*</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.112)</td>
<td>(0.198)</td>
<td>(0.269)</td>
<td>(0.273)</td>
<td>(0.263)</td>
</tr>
<tr>
<td>Return Horizon</td>
<td>6 Weeks</td>
<td>10 Weeks</td>
<td>20 Weeks</td>
<td>30 Weeks</td>
<td>40 Weeks</td>
<td>50 Weeks</td>
</tr>
<tr>
<td>$S_{o,t}$</td>
<td>0.634**</td>
<td>0.821***</td>
<td>0.460</td>
<td>0.665*</td>
<td>0.687</td>
<td>1.827***</td>
</tr>
<tr>
<td></td>
<td>(0.256)</td>
<td>(0.233)</td>
<td>(0.314)</td>
<td>(0.366)</td>
<td>(0.452)</td>
<td>(0.542)</td>
</tr>
</tbody>
</table>

Panel B. Equally Weighted Sentiment Index

<table>
<thead>
<tr>
<th>Return Horizon</th>
<th>1 Day</th>
<th>1 Week</th>
<th>2 Weeks</th>
<th>3 Weeks</th>
<th>4 Weeks</th>
<th>5 Weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{o,t}$</td>
<td>0.094***</td>
<td>0.331***</td>
<td>0.439**</td>
<td>0.614**</td>
<td>0.572**</td>
<td>0.673**</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.115)</td>
<td>(0.202)</td>
<td>(0.273)</td>
<td>(0.273)</td>
<td>(0.267)</td>
</tr>
<tr>
<td>Return Horizon</td>
<td>6 Weeks</td>
<td>10 Weeks</td>
<td>20 Weeks</td>
<td>30 Weeks</td>
<td>40 Weeks</td>
<td>50 Weeks</td>
</tr>
<tr>
<td>$S_{o,t}$</td>
<td>0.840***</td>
<td>1.044***</td>
<td>0.569</td>
<td>1.006**</td>
<td>0.967*</td>
<td>2.307***</td>
</tr>
<tr>
<td></td>
<td>(0.268)</td>
<td>(0.261)</td>
<td>(0.347)</td>
<td>(0.417)</td>
<td>(0.496)</td>
<td>(0.629)</td>
</tr>
</tbody>
</table>

Notes: This table reports the horizon-specific effect of overnight CISI sentiment on the cumulative returns of the Small-Minus-Big portfolio, formed as the difference of the returns of small-stock and big-stock portfolios. These portfolios are value-weighted, monthly rebalanced, and contain the 100 firms with the smallest and the biggest market capitalizations among the 300 constituent stocks in the CSI 300, respectively. The results are organized in the same way as in Table VI.
TABLE VIII
Multi-horizon Regressions for the Growth-Minus-Value Portfolio

<table>
<thead>
<tr>
<th>Panel A. Value-weighted Sentiment Index</th>
<th>Return Horizon</th>
<th>1 Day</th>
<th>1 Week</th>
<th>2 Weeks</th>
<th>3 Weeks</th>
<th>4 Weeks</th>
<th>5 Weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_{o,t} )</td>
<td></td>
<td>0.063***</td>
<td>0.161</td>
<td>0.285</td>
<td>0.369</td>
<td>0.332</td>
<td>0.362</td>
</tr>
<tr>
<td></td>
<td>(0.032)</td>
<td>(0.135)</td>
<td>(0.252)</td>
<td>(0.328)</td>
<td>(0.314)</td>
<td>(0.303)</td>
<td></td>
</tr>
<tr>
<td>Return Horizon</td>
<td></td>
<td>6 Weeks</td>
<td>10 Weeks</td>
<td>20 Weeks</td>
<td>30 Weeks</td>
<td>40 Weeks</td>
<td>1 Year</td>
</tr>
<tr>
<td>( S_{o,t} )</td>
<td></td>
<td>0.503</td>
<td>0.847***</td>
<td>1.120***</td>
<td>1.591***</td>
<td>2.209***</td>
<td>2.953***</td>
</tr>
<tr>
<td></td>
<td>(0.311)</td>
<td>(0.306)</td>
<td>(0.338)</td>
<td>(0.358)</td>
<td>(0.442)</td>
<td>(0.524)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B. Equally Weighted Sentiment Index</th>
<th>Return Horizon</th>
<th>1 Day</th>
<th>1 Week</th>
<th>2 Weeks</th>
<th>3 Weeks</th>
<th>4 Weeks</th>
<th>5 Weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_{o,t} )</td>
<td></td>
<td>0.100***</td>
<td>0.226*</td>
<td>0.348</td>
<td>0.446</td>
<td>0.374</td>
<td>0.382</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(0.130)</td>
<td>(0.243)</td>
<td>(0.304)</td>
<td>(0.287)</td>
<td>(0.275)</td>
<td></td>
</tr>
<tr>
<td>Return Horizon</td>
<td></td>
<td>6 Weeks</td>
<td>10 Weeks</td>
<td>20 Weeks</td>
<td>30 Weeks</td>
<td>40 Weeks</td>
<td>1 Year</td>
</tr>
<tr>
<td>( S_{o,t} )</td>
<td></td>
<td>0.527*</td>
<td>0.877***</td>
<td>0.994***</td>
<td>1.633***</td>
<td>2.278***</td>
<td>2.774***</td>
</tr>
<tr>
<td></td>
<td>(0.280)</td>
<td>(0.319)</td>
<td>(0.355)</td>
<td>(0.385)</td>
<td>(0.482)</td>
<td>(0.574)</td>
<td></td>
</tr>
</tbody>
</table>

**Notes:** This table reports the horizon-specific effect of overnight CISI sentiment on the cumulative returns of the Growth-Minus-Value portfolio, formed as the difference of the returns of growth-stock and value-stock portfolios. These portfolios are value-weighted, monthly rebalanced, and contain the 100 firms with the lowest and the highest earnings-to-price ratios among the 300 constituent stocks in the CSI 300, respectively. The results are organized in the same way as in Table VI.
<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{o,t}$</td>
<td>0.113***</td>
<td>0.127***</td>
<td>0.123***</td>
<td>-0.129</td>
<td>0.129***</td>
</tr>
<tr>
<td>(0.018)</td>
<td>(0.018)</td>
<td>(0.017)</td>
<td>(0.984)</td>
<td>(0.022)</td>
<td></td>
</tr>
<tr>
<td>$S_{o,t} \times A_{o,t}$</td>
<td>0.085***</td>
<td>0.071***</td>
<td>0.067**</td>
<td>0.075***</td>
<td></td>
</tr>
<tr>
<td>(0.018)</td>
<td>(0.018)</td>
<td>(0.028)</td>
<td>(0.020)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_{o,t} \times RV_t$</td>
<td>0.054***</td>
<td>0.053***</td>
<td>0.053***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.017)</td>
<td>(0.019)</td>
<td>(0.017)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_{o,t} \times Vlm_t$</td>
<td>0.011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.043)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_{o,t} \times LowGrowth_t$</td>
<td></td>
<td></td>
<td></td>
<td>-0.017</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.032)</td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Adj-$R^2$(%)</td>
<td>37.6</td>
<td>38.8</td>
<td>39.2</td>
<td>39.2</td>
<td>39.2</td>
</tr>
</tbody>
</table>

Notes: This table reports the estimation results for the specification $R_{o,t+1} = \alpha + \beta \times S_{o,t} + (\gamma_1 A_{o,t} + \gamma_2 RV_t + \gamma_3 Vlm_t + \gamma_4 LowGrowth_t) \times S_{o,t} + \delta \times Controls_t + \epsilon_{t+1}$. The overnight sentiment $S_{o,t}$ is the value-weighted CISI computed using messages during the overnight period between trading days $t$ and $t+1$. The overnight attention $A_{o,t}$ is constructed as the logarithm of the number of messages posted during the overnight period. Both series are normalized to have zero mean and unit standard deviation. $RV_t$ and $Vlm_t$ are logarithms of the daily realized volatility and trading volume of the CSI 300 in day $t$, respectively. The indicator for economic downturn $LowGrowth_t$ is a binary variable that equals to one when the quarterly real GDP growth rate is below the sample median. The control variables include lagged daily returns, log volume, and log realized volatilities of the CSI 300 up to 5 days, the latest daily returns of the S&P 500 index and NYMEX oil futures, and weekday dummies. Standard errors displayed in parentheses are computed using the Newey–West estimator with eight lags; *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from July 1, 2008 to February 14, 2018.
**TABLE X**  
PREDICTIVE REGRESSIONS FOR REALIZED VOLATILITY

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.076***</td>
<td>0.075***</td>
<td>0.076***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.021)</td>
<td>(0.021)</td>
<td>(0.021)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$</td>
<td>S_{o,t}</td>
<td>$</td>
<td>-0.064***</td>
<td>-0.061***</td>
<td>-0.065***</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.015)</td>
<td>(0.016)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A_{o,t}$</td>
<td>0.121***</td>
<td>0.113***</td>
<td>0.067***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td>(0.022)</td>
<td>(0.026)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$</td>
<td>S_{o,t}</td>
<td>\times A_{o,t}$</td>
<td></td>
<td>0.064***</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.019)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_{o,t} \times A_{o,t}$</td>
<td></td>
<td>0.013</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.013)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Adj-$R^2$(%)</td>
<td>74.7</td>
<td>74.7</td>
<td>74.8</td>
<td>75.2</td>
<td>75.3</td>
</tr>
</tbody>
</table>

Notes: This table reports the estimation results for the specification $RV_{t+1} = \alpha + \beta_1 \times |S_{o,t}| + \beta_2 \times S_{o,t} + \beta_3 \times A_{o,t} + \gamma_1 \times |S_{o,t}| \times A_{o,t} + \gamma_2 \times S_{o,t} \times A_{o,t} + \delta \times Controls_t + \epsilon_{t+1}$. The dependent variable $RV_{t+1}$ is the logarithm of the realized volatility of the CSI 300 index constructed as the sum of squared 5-minute returns over day $t + 1$. The overnight sentiment $S_{o,t}$ is the value-weighted CISI computed using messages during the overnight period between trading days $t$ and $t + 1$. The overnight attention $A_{o,t}$ is constructed as the logarithm of the number of messages posted during the overnight period. Both series are normalized to have zero mean and unit standard deviation. The control variables include daily, weekly, and monthly lags of the log realized volatilities ($RV_t$, $(1/5) \sum_{j=0}^4 RV_{t-j}$ and $(1/22) \sum_{j=0}^{21} RV_{t-j}$), and those of the log trading volume ($Vlm_t$, $(1/5) \sum_{j=0}^4 Vlm_{t-j}$ and $(1/22) \sum_{j=0}^{21} Vlm_{t-j}$) of the CSI 300 index, five lagged daily returns of the CSI 300 index, the latest close-to-close returns of the S&P 500 index and NYMEX oil futures, and the weekday dummies. Standard errors displayed in parentheses are computed using the Newey–West estimator with eight lags; *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from July 1, 2008 to February 14, 2018.
### TABLE XI
PREDICTIVE REGRESSIONS FOR TRADING VOLUME

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(</td>
<td>S_{o,t})</td>
<td></td>
<td>0.028***</td>
<td></td>
<td>0.025***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td></td>
<td>(0.008)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(S_{o,t})</td>
<td>0.006</td>
<td></td>
<td>0.008</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.006)</td>
<td></td>
<td>(0.006)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(D_{o,t})</td>
<td>0.010*</td>
<td></td>
<td>0.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.006)</td>
<td></td>
<td>(0.007)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(CEPU_t)</td>
<td>-0.003</td>
<td></td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td></td>
<td>(0.005)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(A_{o,t})</td>
<td></td>
<td>0.051***</td>
<td>0.028**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.009)</td>
<td>(0.012)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(</td>
<td>S_{o,t}</td>
<td>\times A_{o,t})</td>
<td>0.036***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(S_{o,t} \times A_{o,t})</td>
<td>0.007</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(D_{o,t} \times A_{o,t})</td>
<td>0.013***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Adj-(R^2) (%)</td>
<td>90.1</td>
<td>90.0</td>
<td>90.0</td>
<td>90.0</td>
<td>90.2</td>
<td>90.4</td>
</tr>
</tbody>
</table>

**Notes:** This table reports the estimation results for the specification \(Vlm_{t+1} = \alpha + \beta_1 |S_{o,t}| + \beta_2 S_{o,t} + \beta_3 D_{o,t} + \beta_4 \times CEPU_t + \beta_5 \times A_{o,t} + (\gamma_1 |S_{o,t}| + \gamma_2 S_{o,t} + \gamma_3 D_{o,t}) \times A_{o,t} + \delta \times Controls + \epsilon_{t+1}\). \(D_{o,t}\) is the overnight textual disagreement measure and \(CEPU_t\) is the prevailing monthly China EPU index developed using the methodology of Baker et al. (2016). The other variables (including all control variables) are described as in Table X. Standard errors displayed in parentheses are computed using the Newey–West estimator with eight lags; *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The sample period is from July 1, 2008 to February 14, 2018.
## TABLE A.1
### CLASSIFICATION ACCURACY WITH RANDOM SAMPLE-SPLITTING

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Classification Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
</tr>
<tr>
<td>(1) LM–EW</td>
<td>31.15</td>
</tr>
<tr>
<td>(2) LM–TFIDF</td>
<td>32.95</td>
</tr>
<tr>
<td>(3) CSM–EW</td>
<td>70.74</td>
</tr>
<tr>
<td>(4) CSM–TFIDF</td>
<td>75.65</td>
</tr>
<tr>
<td>(5) One-hot SVM</td>
<td>77.40</td>
</tr>
<tr>
<td>(6) Word2vec SVM</td>
<td>81.45</td>
</tr>
<tr>
<td>(7) Word2vec CNN</td>
<td>81.24</td>
</tr>
</tbody>
</table>

Notes: This table shows the out-of-sample classification accuracy for different methods with the labeled sample split into the training (60%), validation (20%) and testing (20%) subsamples randomly. The textual methods include the dictionary method based on the Chinese translation of Loughran and McDonald’s (2011) dictionary (LM), and the Chinese Stock Market (CSM) dictionary, the support vector machine (SVM) with one-hot and word2vec representations, and the convolutional neural network (CNN) based on the word2vec representation. Two weighting schemes are used for each of the dictionary methods: equal weighting (EW) and term frequency inverse document frequency (TFIDF) weighting. The column labeled “All” reports the proportions of correct classification in the testing sample for messages in all three categories {positive, negative, ambiguous}. The “Positive” and “Negative” columns report the classification accuracy within the subsamples of positive and negative messages, respectively.
<table>
<thead>
<tr>
<th>Classifier</th>
<th>All</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) LM–EW</td>
<td>31.62</td>
<td>24.06</td>
<td>25.70</td>
</tr>
<tr>
<td>(2) LM–TFIDF</td>
<td>33.41</td>
<td>26.06</td>
<td>28.46</td>
</tr>
<tr>
<td>(3) CSM–EW</td>
<td>69.65</td>
<td>74.57</td>
<td>72.12</td>
</tr>
<tr>
<td>(4) CSM–TFIDF</td>
<td>73.93</td>
<td>79.20</td>
<td>80.15</td>
</tr>
<tr>
<td>(5) One-hot SVM</td>
<td>74.83</td>
<td>74.91</td>
<td>88.53</td>
</tr>
<tr>
<td>(6) Word2vec SVM</td>
<td>78.68</td>
<td>81.31</td>
<td>88.73</td>
</tr>
<tr>
<td>(7) Word2vec CNN</td>
<td>80.54</td>
<td>82.23</td>
<td>89.05</td>
</tr>
</tbody>
</table>

Notes: This table shows the out-of-sample classification accuracy for different methods with the labeled sample split into the training (60%), validation (20%) and testing (20%) subsamples ordered in time. The textual methods include the dictionary method based on the Chinese translation of Loughran and McDonald’s (2011) dictionary (LM), and the Chinese Stock Market (CSM) dictionary, the support vector machine (SVM) with one-hot and word2vec representations, and the convolutional neural network (CNN) based on the word2vec representation. Two weighting schemes are used for each of the dictionary methods: equal weighting (EW) and term frequency inverse document frequency (TFIDF) weighting. The column labeled “All” reports the proportions of correct classification in the testing sample for messages in all three categories {positive, negative, ambiguous}. The “Positive” and “Negative” columns report the classification accuracy within the subsamples of positive and negative messages, respectively.
FIGURE I
Message Posting Activity within a Week

Notes: This figure plots the average number of messages of each day of the week at half-hour frequency. The sample period is from July 1, 2008 to February 14, 2018.
FIGURE II
Price and Volume of the CSI 300 Index

Notes: This figure shows the daily closing price (solid line) and the trading volume (shaded) of the CSI 300 index. The daily volume (in billion shares) is constructed as the aggregate of the daily volume of the CSI 300 stocks. The daily closing price is the CSI 300 index at 3:00 pm Beijing time.
FIGURE III
Time Series of the Daily CISI

Notes: This figure plots the time series of the daily value-weighted CISI textual sentiment index. The daily sentiment index is formed using messages posted between 3:00 pm on trading day $t - 1$ and 3:00 pm on trading day $t$. The index is normalized to have zero mean and unit standard deviation. The solid line is the 22-day moving average of the daily series.
FIGURE IV
Market Schedule and Conditioning Information Set

<table>
<thead>
<tr>
<th>Conditioning Information Set</th>
<th>Trading Day $t+1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lagged local market information (price, volume, volatility)</td>
<td>Call Auction</td>
</tr>
<tr>
<td>Overnight Textual Messages</td>
<td>B R E A K</td>
</tr>
<tr>
<td>S&amp;P 500 Return (4pm-4pm EST)</td>
<td>Continuous Auction</td>
</tr>
<tr>
<td>NYMEX Oil Futures Return (5pm-5pm EST)</td>
<td></td>
</tr>
</tbody>
</table>

3 PM | 9:15 AM | 9:30 AM
FIGURE V

Time-Varying Effect of Textual Sentiment on Stock Return

Notes: This figure presents the estimated time-varying effect of textual sentiment on stock returns. The state-dependent estimate (solid line) is \( \hat{\beta}_{t}^{Roll} = \hat{\beta} + \hat{\gamma}_1 \times \frac{1}{250} \sum_{j=0}^{249} A_{o,t-j} + \hat{\gamma}_2 \times \frac{1}{250} \sum_{j=0}^{249} RV_{t-j} \), where \( \hat{\beta} \), \( \hat{\gamma}_1 \) and \( \hat{\gamma}_2 \) are obtained from the full sample estimates of the model \( R_{o,t+1} = \alpha + \beta \times S_{o,t} + (\gamma_1 A_{o,t} + \gamma_2 RV_t) \times S_{o,t} + \delta \times Controls_t + \epsilon_{t+1} \). The rolling-window estimate (dashed line) \( \hat{\beta}_{t}^{Roll} \) is obtained by estimating \( R_{o,t+1} = \alpha_t + \hat{\beta}_{t}^{Roll} \times S_{o,t} + \hat{\delta}_{t}^{Roll} \times Controls_t + \epsilon_{t+1} \) within the rolling window \([t-249, t]\); the 95% confidence intervals (shadow area) are computed using the Newey–West standard error with eight lags. The variables are defined in the same way as in Table IX. In all estimations, the control variables include lagged daily CSI 300 returns, log volume, and log realized volatilities up to 5 days, the latest daily returns of the S&P 500 index and NYMEX oil futures, and weekday dummies.