Language and Domain Specificity: A Chinese Financial Sentiment Dictionary[*]

Zijia Du
Shanghai Jiao Tong University

Alan Guoming Huang
University of Waterloo

Russ Wermers
University of Maryland

Wenfeng Wu
Shanghai Jiao Tong University

September 2021

[*] Du is at Shanghai Jiao Tong University, email: shduzijia@sjtu.edu.cn; Huang is at the University of Waterloo, email: aghuang@uwaterloo.ca; Wermers is at the University of Maryland, email: wermers@umd.edu; and Wu is at Shanghai Jiao Tong University, email: wfwu@sjtu.edu.cn. Joey Bao provides excellent research assistance. The authors gratefully acknowledge helpful comments from two anonymous referees, Fabrice Schloegel, Grant Wang, Yongxiang Wang, Paul Winter, Baozhong Yang (discussant), Chen Yang (discussant), Jiaquan Yao, Jiaxing You, Bohui Zhang, and seminar and conference participants at AGFiq Asset Management, China Financial Research Conference 2021, China International Risk Forum 2021, and UBS Quant Insight Series. Wenfeng Wu acknowledges financial support from the National Science Fund Committee of China (Nos. 71850010; 71772121; 71790592). All errors are ours.

Language and Domain Specificity: A Chinese Financial Sentiment Dictionary

Abstract

We use supervised machine learning to develop a financial sentiment dictionary from 3.1 million Chinese-language financial news articles. Our dictionary maps semantically similar words to a subset of human-expert generated financial sentiment words. In validation tests, our dictionary scores the sentiment of articles consistently with human reading of full articles. In return association tests, our dictionary outperforms and subsumes previous Chinese financial sentiment dictionaries, such as direct translations of Loughran and McDonald's (2011) English-language financial dictionary. We also generate a list of politically-related positive words that is unique to China; we find that this list has a weaker association with returns than does the list of other positive words. We demonstrate that state media uses more politically-related positive and fewer negative words, and exhibits a sentiment bias. This bias renders the state media's sentiment as less return-informative. Our findings demonstrate that dictionary-based sentiment analysis exhibits strong language and domain specificity.

1. Introduction

One might have fond memories of the 2003 American film *Lost in Translation*, where the starring actor, Bill Murray, was baffled in communicating with a Tokyo director because subtle meanings were lost in translation. Languages have delicate linguistics and connotations that are often difficult to capture in translation—for example, there is no unique Chinese equivalent of the word "yes."[1] As an example, our translation to Chinese of the defacto English finance sentiment-word dictionary of Loughran and McDonald (2011) differs significantly from those published by others.[2] In fact, recent discussions of a potential universal "common financial language" evolve into a call to use "applied linguistics" to accommodate for "richness and depth" of languages (Robinson, 2018). That is, in an applied field such as financial economics, cross-language work must consider cultural, societal, geographic, and regulatory subtleties among languages.

In this paper, rather than using a simple translation of an existing English financial dictionary, we build, from the ground-up, a Chinese financial sentiment word dictionary from processing over three million Chinese-language financial news articles.[3] In doing so, we note that the expression of fondness (which is related to sentiment), or the lack thereof, carries perhaps the greatest subtleties in a language. Chinese is among the most subtle of languages, due to its relatively small number of characters and its large number of words or expressions formed out of the many combinations of these characters.[4] Accordingly, it is important—in building a comprehensive and precise financial sentiment dictionary—to start from the basics of the Chinese language, rather than attempting to translate English-language financial dictionaries to the Chinese cultural and language setting.

The emphasis on the cultural and societal context of cross-language work has a parallel with

[1] Depending on context, a "yes" can correspond to the Chinese equivalents of "right", "correct", "okay", "fine", "good", "no problem", "have done so", and "can", among others.

[2] For example, You, Zhang, and Zhang (2018; YZZ) publish a Chinese translation of the Loughran and McDonald (2011; LM) dictionary, augmented by a manual collection of sentiment words from financial news. Our own translation of LM yields only about a 20% overlap with YZZ.

[3] By financial sentiment words, we refer in this paper to words that carry a negative or positive connotation in a financial context.

[4] For example, it is well known that the mere knowledge of 900 Chinese characters enables a person to read 90% of newspapers. In English, Loughran and McDonald (2011), among others, show that a general dictionary such as the Harvard Psychosociological Dictionary is unfit for sentiment analysis, as it, for instance, misclassifies negative tone roughly 75% of the time when examining annual reports.

the literature that relates culture to the general fields of economics and finance (e.g., La Porta et al., 1998; Stulz and Williamson, 2003; De Jong, 2009). Cultural and societal norms impact the economic "values" held by a society (De Jong, 2009). In the context of financial sentiment, for instance, Bradshaw, Huang, and Tan (2019) document that financial analysts' degree of optimism is related to the "well-being" of a country's institutional infrastructure. In China, the media is largely controlled by the state. "*Mind Politics*," a literal translation of the Chinese slogan to emphasize coherence with central government guidelines, has become a norm in contemporaneous culture and has been pervasive in the Chinese media. Such Mind Politics words are rarely negative, and are unique to Chinese, as their counterparts barely appear in English financial news in countries such as the U.S. and the UK. Apart from the usual negative and positive sentiment word list, we identify, from our corpus of financial news, a list of Mind Politics words that are presumably politically-driven to be positive ("political words").

We use supervised machine learning of "word embeddings" to develop our dictionary. We rely on a recent computational linguistic tool of Word2vec (e.g., Mikolov et al., 2013; Jurafsky and Martin, 2019), which employs a neural network algorithm to find semantically similar words to a given target word in a large body of text. Word2vec matches each word to a vector of real numbers that contains the semantic and syntactic information (henceforth, "word embeddings", e.g., Loughran and McDonald, 2011). The algorithm then identifies words that share the highest degrees of common contexts as "semantically close pairs" by maximizing their vector cosine similarity.

We retrieve all corporate news articles for the period from January 2013 to August 2019 on finance.sina.com.cn, the most visited Chinese finance website that streams real-time news and stock information, totaling 3.1 million articles. To create a starting list of target words, we manually read 2,500 randomly selected articles, and pick "seed words" that are positive, negative, or political by their sentiment indicators. We train the Word2vec model through four iterations, sequentially incremental in corpus size (the number of news articles) and in starting seed word size. To incorporate the current literature, we augment the Word2vec training with all sentiment words from YZZ as another round of "seeding."[5] In each of these rounds, we intervene through independent reviews of the output by humans (finance graduate students and ourselves), and filter out the output that we deem unsuitable. As a comparison, for their entire list of sentiment words, YZZ manually read 2,000 news articles, and Yan et al. (2019) read 24 IPO prospectuses. In aggregate, our human reviews remove 80% of the words output by Word2vec. The human "supervision" during the machine learning iterations ensures that the output is consistent with Chinese language "field expertise."

We benchmark our dictionary against three alternative dictionaries: our translation of Loughran and McDonald (2011), augmented with appropriate synonyms ("LM dictionary"); the YZZ dictionary; and the intersection of three general sentiment dictionaries discussed in Huang, Wu, and Yu (2019) ("Generic dictionary"). In total, we identify 6,660 negative, positive, and political words, larger than each of these three dictionaries. Moreover, our dictionary is distinctive from existing dictionaries, as only 39% of our words are covered in these other dictionaries, combined. More specifically, 48% of our negative words are covered by the union of these other dictionaries, while only 19% of our political words are covered (mostly in the YZZ dictionary).

To validate our dictionary, we apply it to firm-specific news to examine the determinants of sentiment, as well as the consistency of article-level sentiment between dictionary word-counting and other means of article sentiment measurement. We count the occurrences of the sentiment words as the news tone, and use the negative, positive, and net negative tone of the news (the frequency of negative minus positive words) as our sentiment measures. We find that smaller and riskier firms, and firms under distress (as proxied by book-to-market ratio), tend to have more negative news, whereas firms with a greater earnings surprise tend to have more positive news—indicating that our sentiment measures are consistent with those documented using U.S. news dictionaries on U.S. media releases (e.g., Tetlock et al., 2008).

To examine the consistency of article-level sentiment of dictionary word-counting, we adopt three methods. First, we manually classify 5,000 news articles as either positive or negative by reading the full articles (without dictionary word counting). We compare dictionary word

[5] Adding the YZZ seed words only marginally increases the size of our dictionary.

counting with the human reading outcome. Second, we break these 5,000 articles into a training set and a test set, apply the more traditional machine learning classifier of support vector machine (SVM) to predict the sentiment of the test set, and examine the consistency between dictionary-word counting and the SVM prediction for the test set. Finally, we download 50,000 articles from the Wind Terminal (the Chinese equivalent of Bloomberg terminal) that the Terminal labels as either positive or negative. Judging the news sentiment by our dictionary-word counting offers close to a 90% matching rate with these three alternative approaches that are much more laborious (and, potentially, error-prone), and the 90% matching rate compares favorably with the literature (e.g., Huang et al. (2014) report SVM accuracy rates of 75-78% for Chinese social media posts). Comparatively, dictionary-word counting is straightforward and efficient to implement, and offers a continuous measure for sentiment—unlike a potentially binary measurement by other approaches.

Next, we find that our sentiment measures are strongly related to stock returns. In multivariate regressions where we control for the "China-4" factors of market, size, value, and turnover used in Liu, Stambaugh, and Yuan (2019), we find that industry- and size-adjusted abnormal returns are statistically significantly related to our sentiment measures from days [-20] through day [1], with the most significant relation taking place on days [-1] and [0]. Measuring the economic magnitude of these relations, however, reveals that only days [-1] and [0] have a return impact that appears to be tradable; a one standard-deviation increase in the net negative tone results in an impact on returns of 51 bps on day [0] and 31 bps on day [-1].[6] None of the days outside [-1, 0] has economic significance exceeding 10 bps, with many being close to zero. Distinctive in China, positive news tone has a higher impact on returns than does negative news tone, contrary to the findings in U.S. markets, where positive news tone is either insignificant or less significant than negative tone in predicting returns (e.g., Tetlock et al., 2008; Huang, Tan, and Wermers, 2020). We also adjust for potential news persistence, but continue to find economic significance for the return effect of sentiment on days [-1, 0]. The day [-1] return association, thus, points to the existence of a limited degree of news leakage in China, consistent with the conventional

[6] Direct transaction costs (including stamp tax and broker commissions) for a round-trip trade are about 20 basis points (bps) during our sample period.

wisdom and the literature (e.g., Li et al., 2017).

We also find that the return associations of our dictionary dominate those of the three alternative dictionaries (LM, YZZ, and the Generic Chinese dictionaries). When the sentiment measures from all these dictionaries are pooled together in return regressions, we find that our measures completely subsume and even reverse those from the LM and Generic dictionaries, and significantly reduce the significance of the YZZ measures. The economic and statistical significance of our measures remain largely intact after controlling for the sentiment measures of these alternative dictionaries. To account for the fact that our full dictionary encompasses the YZZ dictionary, we also use a stripped version of our dictionary that we derive in the Word2vec training before adding the YZZ seed words. In doing so, we still find that the return impact of the YZZ dictionary sentiment is much smaller than ours. YZZ and our dictionary do share a common ordinal sentiment ranking for the majority of news articles; and for these articles our net-negative tone measure largely subsumes the YZZ measure for these news articles.

Lastly, and as an important contribution of our paper, we show that, while our list of political words has a weaker implication for returns, it has the potential to serve as an indicator for sentiment bias in Chinese news. Recent literature proposes to measure media bias in China based on the coverage of "party-line" content, such as the number of mentions of party leaders (e.g., Qin**,** Strömberg, and Wu, 2018), and shows that state media (relative to business or market-oriented media outlets) showcases its political bias by more frequent mentions of political nouns (e.g., Piotroski, Wong, and Zhang, 2017). Additionally, the literature reports that state media issues fewer negative corporate news articles and that news stories by state media have lower value relevance (e.g., Piotroski, Wong, and Zhang, 2017; YZZ). Using our more precise quantification of word sentiment, we show that the state media exhibits a "sentiment bias" by using more politically-inclined positive words and fewer negative words (i.e., in our dictionary), and that stock returns are less sensitive to this sentiment bias in state media releases. Importantly, state media's sentiment bias is distinct from previously documented "media bias" embodied in the mentions of political entities or nouns. Given the ubiquity of politically inclined words in modern Chinese news, our results highlight the necessity to isolate the list of political words when studying the news sentiment in China.

To summarize, our main contribution is that we create a dictionary of financial sentiment words for China based solely on financial news in Chinese. The current practice of Chinese financial sentiment words is heavily based on Chinese translations of Loughran and McDonald (2011) lexicons (e.g., Wang and Wu, 2015; Xie and Lin, 2015; Lin and Xie, 2016; YZZ; Zeng, Zhou, and Zhang, 2018; Li et al., 2019, to name a few). The power of such a simple translation is dubious, as sentiment words are both domain and language specific, and do not render themselves to a simple and mechanical translation from English. As further evidence, our paper demonstrates that a simple translation of the LM English dictionary is subsumed and dominated by our dictionary in sentiment-return associations. Authors attempting to ameliorate this problem often include additional words from manually reading financial texts (e.g., YZZ; Zhou, Zhang, and Zeng, 2019). We differ from this strand of literature by starting afresh, without imposing the constraints imposed by using a forced Chinese translation of the English-language LM dictionary, and by using the widest possible set of financial news from credible sources that we can practically cover.

In an earlier study in machine learning, Mao et al. (2014) propose a fully automated approach to creating a Chinese sentiment dictionary, based on the return reaction following sentiment words in news headlines expanded from a set of seed words. Yao et al. (2021), use a deep learning technique called “long short-term memory” to construct sentiment lexicons of Chinese words based on three-day stock returns following firms’ annual reports and social media posts. In our supervised machine learning process, we ensure that “seed words must be accurate and semantically unambiguous” (Mao et al., 2014, p. 6), so that we similarly avoid potential errors and ambiguity that could propagate in the expanded lexicon. Different from these authors, we supervise the output of machine learning by human expertise; and stock returns are used, instead, to discipline our approach and test the efficacy of our dictionary. We also add to the recent literature utilizing the word embedding approach in various settings of finance (e.g., Hanley and Hoberg, 2019; Cong, Liang, and Zhang, 2020). Similar to us, Li, Mai, Shen, and Yan (2021) use a supervised Word2vec method to generate a list of dictionary words for corporate culture. Our paper joins this stream of literature in showing that word embedding is an effective approach in dealing with financial texts.

Importantly, we also are the first to create a list of politically inclined positive words that is separate from a generally positive word list. We show that political words are important in indicating the Chinese media's sentiment bias. Benefits of our dictionary-based approach include that it is straightforward to implement and potentially applicable to other texts (such as annual reports), and that the process is explainable, in terms of not being a "black-box," when compared with many other machine-learning methods. Overall, we highlight the importance of a language and domain-specific dictionary for China's stock markets. As China now ranks as the second largest stock market in the world through having two of the ten largest stock exchanges in the world, we believe that a suitable sentiment dictionary for financial texts is of significant economic importance.[7]

2. Dictionary Construction and Firm-Specific News Sentiment

2.1 The firm news sample

To construct our dictionary, we first obtain a large financial word corpus from native Chinese-language news sources. Specifically, we retrieve news, as of August 2019, from finance.sina.com.cn, the most visited China-domiciled finance website that streams real-time news and stock information in Mandarin.[8] We download all "Firm News," dated as far back as possible on the website, for all A-share listed stocks on the Shanghai Stock Exchange and Shenzhen Stock Exchange, two major stock exchanges in mainland China.[9] Firm News includes all news pertinent to a corporation, such as firm-originated announcements, press coverage, and market commentary, but excludes the firm's public regulatory filings.

We download 3,078,175 news articles for the period from January 2013 to August 2019, for 3,557 firms in total. A large quantity of news articles appears on the website, starting November 2013. Figure 1 plots the number of news articles by month. Each month has, typically, greater than 20,000 articles, with the years 2014 and 2015 generally having more observations than other years.

[7] See, e.g., https://en.wikipedia.org/wiki/List_of_stock_exchanges.

[8] See, e.g., the Alexa ranking of business websites (http://chinanetrank.com/index.php?c=domain&a=type&tid=44) or Global Chinese Website Rankings (http://www.cwrank.com/main/rank.php?subtype=money&page=1).

[9] Both exchanges have another class of listed shares called B-shares, which are denominated and traded in U.S. or Hong Kong dollars. The B-share market is tiny and illiquid; and, as a vintage market, it has not listed any new IPOs since 2001.

The number of news articles across time is mostly consistent with the fortunes of the Wind All-A market index,[10] which is also plotted in Figure 1. Overall, there is more media coverage of stocks during market upswings and booms. The firms covered in our news-articles sample represent the vast majority of the firm population in the market; for example, there were 3,565 listed stocks at the end of 2018 on the two stock exchanges.[11]

[Insert Figure 1 about here.]

All news articles are sourced from an established media outlet and, excluding firm-initiated announcements, almost always have author or editor names. We compare the Firm News on finance.sina.com.cn with that from the Wind Terminal for 10 randomly selected firms over our sample period, and find that 85% of the articles in the former are identically sourced in the latter.[12] Articles on Sina and Wind are from similar sources of mainstream media; the Internet Appendix to this paper provides the list of 64 news sources of our articles. 45.4% of our news articles are sourced from four media outlets designated by the China Securities Regulatory Commission (the Chinese equivalent of the U.S. SEC) through which firms may publish public announcements and regulatory filings.[13] We believe that our news sample, therefore, is representative of the general supply of news about most listed mainland Chinese firms to the market.

2.2 A Chinese dictionary through supervised learning by Word2vec

We use supervised machine learning to generate a Chinese financial sentiment word dictionary from the over three million downloaded articles. The computational linguistic tool that we use is Word2vec (e.g., Mikolov, Corrado, Chen, and Dean, 2013; Mikolov et al., 2013; and Jurafsky and Martin, 2019), which employs a neural network algorithm to find semantically similar words to a given target word in a large corpus (body of text). After parsing the corpus into words

[10] The index is compiled by Wind Information Co., Ltd., the largest financial data provider to professionals in China, to represent the performance of the overall mainland Chinese stock market.

[11] Delisting in the A-share market is rare during our sample period; for example, there were only five stocks delisted in 2018.

[12] The news data on Wind is not licensed to us, and it would, otherwise, require manual downloading, had it been granted to us. Nevertheless, this statistic indicates that our Firm News is comprehensive.

[13] These four designated news outlets are China Securities Journal, Shanghai Securities News, Securities Times, and Securities Daily. Firms can also publish announcements on the websites of the two stock exchanges, as well as Juchao Information's website.

or phrases (tokens), Word2vec maps the tokens to vectors of real numbers (word embeddings) to arrive at a vector space, where each unique token takes a specific vector value. It then uses neural network algorithms (such as gradient descent) to group similar vectors to arrive at semantically close words. The benchmark of judging vector similarity is usually cosine similarity, which measures the "closeness" of two vectors. For a given target word, the algorithm collects its nearby words as the "positive" sample, and randomly chooses some words outside of the positive sample as the "negative" sample. The positive sample serves as the context of the target word. Word2vec iteratively maximizes the similarity of the target word and the context words from positive samples while minimizing the similarity of the target word and the context words from negative samples, such that words that share the highest degrees of common contexts are placed close to one another to arrive at semantically close pairs.

To use Word2vec, we need a starting dictionary as the "training" sample. As there does not exist an established Chinese financial sentiment dictionary, we first implement our methodology with a starting sentiment list of seed words from reading 500 randomly selected, 100-word-minimum news articles evenly distributed across 10 industries. Following Huang, Wu, and Yu (2019), we use Jieba, a widely-used Chinese text segmentation Python package, to tokenize news articles.[14] We employ four graduating and graduate Chinese finance-major students to each independently read the news articles, supervised by a doctoral student and ourselves. We pick a particularly positive, negative, or politically-inclined positive ("political") word only if all of the team members converge on the word.[15] The first three rows of Panel A of Table I summarize that we iteratively read the 500 articles over three rounds to identify starting lists of words. We start with 50 articles in the first round, supplemented by 250 and 200 additional articles, respectively,

[14] Tokenization breaks text into word units. We compare the following two tokenization packages: i) Stanford Word Segmenter developed by Stanford University, and ii) NLPIR-ICTCLAS developed by the Chinese Academy of Sciences, and find that Jieba segments Chinese financial news more accurately. See the Internet Appendix for examples of this comparison.

[15] Political words are usually in two or four characters that often feature idiom-like slogans. For example, literally, the word "Zhong Zhi Cheng Cheng" means "build a city by uniting together," "Yu Shi Ju Jin" means "progress with times," and "Shang Xia Yi Xin" means "everybody top and bottom unite together."

over the next two rounds. We converge to 334 negative words, 249 positive words, and 122 political words as our starting dictionary (the sum of the first three rows in Panel A).[16]

[Insert Table I about here.]

We next use this starting dictionary as an input to Word2vec. We train the Word2vec model for the starting dictionary through three iterations. Since the Word2vec model simply outputs computationally close words, we employ human expertise to review the output of every iteration to ensure that the words chosen by Word2vec are indeed semantically close in our domain of financial news articles. The first three rows of Panel B of Table I show the additional sentiment words in each of the three iterations. In Iteration 1, we choose 100 large firms (576,153 articles), and Word2vec outputs an additional 1,858 negative, 1,730 positive, and 878 political words to our starting dictionary of Panel A. After independent human reviews by the team members of the Word2vec word output, we converge to about a third of these words, i.e., 594, 506, and 337, respectively (row 1 in Panel B). We then add, to our starting dictionary, this new converged word list before starting Iteration 2, which expands the sample to 1,000 firms and 1.8 million articles. The third round encompasses all firms and articles (row 3 in Panel B). While we iteratively add the identified synonyms, we recycle the news articles, as the used corpus may contain synonyms to newly added words in a later iteration.

The exercise above illustrates the feasibility of generating the dictionary through our supervised machine learning method. To produce a robust dictionary, it is necessary to use a diverse set of seed words. To this end, we augment the above process with two sets of seed words: (i) seed words from reading another 2,000 randomly selected articles across different industries, and (ii) additional sentiment words from the YZZ dictionary. YZZ translate Loughran and McDonald’s (2011; LM) positive and negative English words, and manually augment the translation by positive and negative words from reading 2,000 financial news articles. In our first augmentation, we do not use the YZZ dictionary but follow their lead of reading 2,000 articles.

Table I shows that, while the manual reading of another 2,000 articles yields incrementally about the same number of sentiment words as our original 500 articles (Round 4, as compared with

[16] We remove sentiment words that are unrelated to economics, business, and finance, such as those used to describe natural sceneries.

Rounds 1 to 3 in Panel A), it results in a significantly smaller number of valid Word2vec synonyms (Iteration 4 in Panel B). Finally, we include all sentiment words from the YZZ dictionary that are not yet covered by the seed words from our reading of 2,500 articles and the four iterations of Word2vec exercises, as our Round 5 seed words. Using these additional YZZ seed words, we produce a very small set of additional valid Word2vec synonyms (Iteration 5 in Panel B).[17] Overall, human reviews filter out 80.3% of the Word2vec output, with an increasingly lower survival rate for Word2vec output in further iterations. We take comfort in the fact that, as we move to further iterations, we incrementally identify fewer valid synonyms despite a much-increased corpus size.

One salient outcome from Table I is that, albeit manual reading of the additional 2,000 news articles and including the YZZ dictionary significantly increase the number of seed words, the number of valid Word2vec synonyms additionally produced from them is small. In fact, 87.7% of all valid synonyms are produced by seed words from reading the first 500 articles. We believe that this is due to the fact that financial news tends to be homogenous in both content and writing, such that a smaller sample is largely representative. In fact, our results remain robust to using the dictionary created with the first 500 articles. Nonetheless, we refer to the aggregate set of sentiment words produced in Panels A and B of Table I as our dictionary.

2.3 Comparison with existing dictionaries

Panel A of Table II summarizes the number of sentiment words in our final dictionary. We produce 2,986 negative words, 2,235 positive words, and 1,439 political words. More than half of the sentiment words are in two Chinese characters, consistent with the fact that two-character words are the dominant form of Chinese words.

[Insert Table II about here.]

Panels B and C of Table II provide a comparison of our dictionary with existing Chinese dictionaries. We compare with three existing dictionaries. First, we follow the literature to use the direct Chinese translation of LM's dictionary of negative and positive words (see, e.g., YZZ; Zhou,

[17] We remove words from the YZZ dictionary that are not treated as a "token" in the Jieba segmenter, and partition their positive words to positive and politically-inclined words. We find that the vast majority of the words in the YZZ dictionary are already covered by our earlier rounds of Word2vec.

Zhang, and Zeng, 2019; Zhong, Dong, and Chen, 2020).[18] A direct translation by us yields 823 negative and 154 positive words.[19] As the direct translation appears small in size, we expand the list by finding the appropriate synonyms of the translated words.[20] We use our expanded Loughran and McDonald (2011) dictionary as our English-origin benchmark (the "LM dictionary"). The second benchmark that we use to compare is the YZZ dictionary. The third benchmark dictionary is the set of common words shared by three general (i.e., not finance-specific) Chinese sentiment-word dictionaries that are widely used in common settings, by Dalian University of Technology, China National Knowledge Infrastructure (commonly known as HowNet), and National Taiwan University, respectively. These three dictionaries each have 8,700 to 21,900 sentiment words; however, there is little overlap of words among the three. In a review paper, Huang, Wu, and Yu (2019) use the intersection set of the three dictionaries for a pilot study of Chinese financial news. We follow these authors and use this common set as our Generic sentiment dictionary. The Generic dictionary has 566 negative and 639 positive words.

Panel B of Table II shows that, with 6,660 words in total, the size of our dictionary is much larger than each of the three existing dictionaries, and larger than the union of these three prior dictionaries. Note that the LM and YZZ dictionaries have about the same number of negative words, yet differ substantially in the number of positive words. This points to one of the major differences in Chinese news—that there exists a large number of unique politically-inclined positive words that are not covered in English (imprecisely captured by the YZZ dictionary, and not captured at all by our LM dictionary). While we include the sentiment words from YZZ, our dictionary is about twice the size of the YZZ dictionary.

What is more distinctive about our dictionary is that we have only a limited degree of overlap with the existing dictionaries. Panel C shows that, among the three categories of sentiment,

[18] Bill McDonald provides the latest version of their dictionary on the website https://sraf.nd.edu/.

[19] The number of translated words using LM is significantly smaller than the number in the original LM English word list, since there is no grammatical tense in Chinese, and words of different forms possessing the same word root in English are often translated into the same word. For example, Loughran and McDonald's (2011) dictionary treats "abandon" and "abandoned" as two distinctive negative words, yet, in Chinese, they are translated to the same Chinese word.

[20] We use Synonyms, a Python package which provides Chinese synonyms for natural language processing. We manually check the list and remove inappropriate words.

our negative-word list is most frequently covered by another dictionary; however, the highest level of overlap is still only 38.35%—with the YZZ dictionary (the positive-word overlap between ours and YZZ is slightly smaller). 72% (= 2,165/3,003) of the words in the YZZ dictionary are covered by us, and those not covered are largely words that are not treated as tokens in Jieba. In untabulated results, we re-tokenize the uncovered YZZ words and find that about one-third of them are, in fact, covered in our dictionary. The degree of overlap between our words and other dictionaries is significantly smaller, and is often in single-digit percentage points. Overall, only 39.4% of our words are covered in these other dictionaries combined—negative words have the highest ratio of 48.0%, followed by 40.7% in positive words, and 19.5% in political words. In sum, Table II suggests that our dictionary is highly distinctive from existing dictionaries, and that our additional list of politically-inclined positive words is only lightly covered by the prior literature.

2.4 Firm-specific news filtering and sentiment measures

After constructing our Chinese dictionary, we proceed by filtering to obtain firm-specific news for cross-sectional news analyses. We observe that a significant number of news articles are related to the broader market or the industry group that the firm belongs to, rather than to the firm tagged by the website. For example, an article entitled "Daily Collection of Firm Announcements" summarizes announcements by multiple firms—usually by simply repeating the announcement headline—but the article is tagged to all of the firms in the article. Although this article helps in constructing our dictionary (where a larger corpus is desirable), assigning the article to the firms mentioned therein would bias against finding firm-specific results, as the article is better suited for more general market sentiment analysis.

In general, we identify a list of key phrases in news headlines that are effective in filtering out market- or industry-related news.[21] After filtering out these news articles, we follow Huang, Tan, and Wermers (2020, hereafter "HTW") to count the frequencies of all mentions of firm identities in assigning the news to a given firm. Specifically, for an article to be confirmed as firm-specific to the firm tagged by our source data, we count the frequency of mentions of each of the

[21] These key phrases include "Daily Collection of Firm Announcements", "Collection of Investment Opportunities", "Industry Research", "Collection of Stocks Hitting Daily Limits Today", "Stocks Sought After by Funds", "Highlights of Important Market News", "Industry Information Revealed", and "League Table of Stock Market".

top three firms. We require that: (i) the tagged firm is among the top three-mentioned firms; and (ii) if the firm is not the most mentioned, the frequency of its mentions is at least 50% of that of the top mention. Condition (i) results in the biggest loss of observations, by about half; however, this condition ensures that our sample of news is, in fact, firm-specific. For condition (ii), the assigned firm appears as the most-mentioned firm for 79.7% of news articles. We also remove news articles that have less than 50 words. After these steps, we are left with 967,919 news articles for 3,552 firms. Appendix A provides further details on the news filtering process.

News can also be stale and repeat a previous news article (e.g., Tetlock, 2011). We follow Fedyk and Hodson (2020) in identifying news "reprints"—news articles that largely repeat a previous article. For each news article that is assigned to a given firm, we examine the headlines of the same-firm news articles during the surrounding [-3, 3] trading days. If the headlines of any two same-firm news articles have an overlap of at least 70% (after excluding stop words and firm names), we treat the later article as a reprint of the earlier article. This process results in an original news article, followed by, if any, its reprints. 12.57% of the articles are identified as reprints. In untabulated results, we find that about two-thirds of the reprint articles are from a media outlet different from that of the original article; among this subset of articles, the headline overlap ratio is 90%. These results suggest that media outlets in China often repeat the same news with little editorial revision. We drop these reprints and keep only the original news as our primary sample.

We calculate the sentiment of the news following the literature (e.g., Tetlock, 2007; Tetlock, Saar-Tsechansky, and Macskassy, 2008; HTW). To reflect the literature's emphasis on negative tone (e.g., Tetlock, Saar-Tsechansky, and Macskassy, 2008), our primary sentiment measure is the net negative tone (*Neg_net*), defined as the number of negative-word occurrences minus positive-word occurrences, divided by the total number of words.[22] We also consider the two components of *Neg_net*: *Neg* (*Pos*), the ratio of negative (positive) word count to the total number of words in the news article. Lastly, we calculate the ratio of political words (*PoliticalPos*). Analogously, we add the dictionary suffix to sentiment measured using other dictionaries; for

[22] We remove stop words from the corpus when counting the total number of words. We identify stop words by aggregating four common lists (see, e.g., https://github.com/goto456/stopwords).

example, the net negative tone by the LM, YZZ, and Generic dictionaries is denoted as, respectively, *Neg_net_LM*, *Neg_net_YZZ*, and *Neg_net_generic*. Appendix B provides the definitions of the variables used in this paper.

2.5 Validation of news sentiment at the firm-day level

We now provide validity checks of news sentiment constructed from our dictionary. We take the mean for each news sentiment measure by trading day for each firm, weighted by the total number of words in each news article during the day. For a given trading day *t*, news articles released after the market close of trading day *t*−1 and before the market close of trading day *t* are combined as being news for day *t*; in subsequent sections, we refer to this definition of day *t* as our event day 0.[23]

A number of noteworthy institutional features could lead to excessive or singular-toned news coverage that could bias our results, in particular, when testing the efficacy of our dictionary on stock returns. As IPO issuance is a heavily regulated process in China in our sample period, newly IPOed stocks tend to receive disproportionally high media attention. Further, stocks that have deemed abnormalities (usually those incurring net losses for two consecutive years) will receive a designated ST (special treatment) status. And, stocks experiencing major events (such as acquisitions of major assets) will likely have a trading halt lasting weeks to months. Due to the existence of daily price change limits, these stocks frequently exhibit a string of hitting price change limits over a period of time (positive 10% of returns in IPOs and most major events, and negative 5% of returns in ST). We, accordingly, remove the first 20 days post the IPO date, all stock-days designated as ST, and the trading halt days plus the first trading day after the halt if the firm experiences more than one day of trading halt. Our final sample consists of 424,758 trading days for 3,539 firms.

Panel A of Table III reports the summary statistics of our sentiment measures for the final sample. The mean of *Neg_net* is negative, indicating that the news tone is, on average, positive. This is likely due to the fact that about half of our news is firm-initiated (see Internet Appendix).

[23] The trading hours of Chinese A-Shares consist of a morning session of 9:30 to 11:30 and an afternoon session of 1:00-3:00 local (Beijing) time.

The literature has documented that U.S. corporate management tends to withhold negative news and exhibits biased optimism in firm announcements (e.g., Kothari, Shu, and Wysocki, 2009). Consistently, YZZ manually categorize the article tone (by reading each entire article) of 200,000 Chinese financial news articles, and find that the average article tone is significantly positive. The mean of *PoliticalPos* (0.0300) is larger than the magnitude as that of *Neg* (0.0276); that is, political sentiment words are more commonplace than negative words. Due to their smaller numbers of words, the measured level of negative and positive sentiment by the LM dictionary (*Neg_LM* and *Pos_LM*) and the Generic dictionary (*Neg_generic* and *Pos_generic*) tend to be smaller than *Neg* and *Pos*, respectively.

[Insert Table III about here.]

Panel B of Table III presents the sample correlations of the sentiment measures. A number of observations are in order. First, the correlation between *Neg_net* and *Neg_net_LM* (*Neg_net_YZZ*) is significant at 0.60 (0.83); and the correlations for the negative and positive legs between our dictionary and the LM (YZZ) dictionary are similar in magnitude. The correlations of our sentiment measures with those of the Generic dictionary are much smaller. Second, *Pos* and *PoliticalPos* are only moderately correlated at 0.35. Thus, although political words are almost always positive by nature in China, the political sentiment differs substantially from non-political sentiment in corporate news. Lastly, the correlation coefficient between *PoliticalPos* and *Pos_YZZ* is 0.58; although only 14.45% of our political words appear in the YZZ dictionary (Table II), this correlation suggests that a significant portion of YZZ's positive sentiment overlaps with our spectrum of political sentiment. In contrast, *PoliticalPos* is modestly related to the positive ratio in other dictionaries. Overall, these results suggest that YZZ is the closest to us. In Section 3, we run a horse race among these dictionaries, and show that our dictionary has the potential to subsume that of YZZ in return implications.

Panel A of Table IV examines the tone trend, by year. Generally, in market upswings, such as 2014 to the first half of 2015 (see Figure 1), the tone is more positive and less negative. The most negative year is 2016, where the market experiences three historical crashes, starting from the second half of 2015 and lasting into the first half of 2016.[24] Panel B of Table IV shows the average tone across the 23 industries in the A-share market. Consistent with the market-wide results, declining industries, such as apparel and mining, have the highest net-negative tones, while rising industries, such as software and information services, catering, and entertainment, enjoy the lowest net-negative tones. Note, also, that the level of politically positive toned news is similar across industries, consistent with a "pure spin" (i.e., noise) that is applied to news across the spectrum.

[Insert Table IV about here.]

As a firm level validity check, in Table V, we regress the sentiment measures on a number of firm and news attributes. The firm attributes include: (i) a Chinese version of the Fama-French (1993) three factors of firm beta, market size, and book-to-market ratio; (ii) past stock price attributes of turnover, return standard deviation, industry- and size-adjusted excess returns of month [–12, –2], and of days [–10, –6], [–5, –3], [–2], and [–1]; and (iii) additional stock attributes of earnings surprise (SUE), dividend yield, firm age since IPO, and whether the firm belongs to the CSI 300 index (an index consisting of 300 blue-chip Chinese stocks) and to a state-owned-enterprise (SOE). The news attributes include the historical number of articles during the past year, and the number of news articles on the given trading day. Except for the news data, the market and financial data are calculated from CSMAR (China Stock Market & Accounting Research) Database. All of the control variables are measured prior to the news day.

[Insert Table V about here.]

Table V shows that firms that are smaller, have higher betas, or have higher book-to-market ratios tend to exhibit a more negative news tone, and that firms with greater earnings surprise tend to exhibit a more positive tone. These findings are consistent with riskier and less profitable firms having more negative news. Further, higher volatility and turnover both lead to less (net) negative news, but this effect primarily lies on the positive side of the tone: both variables drive *Pos* much more significantly than they do *Neg*. This is most likely due to the fact that the Chinese market often features rampant speculation—market sentiment can become more "exuberant" in the

[24] The Chinese press dubs these three crashes as Market Crash 1.0, 2.0 and 3.0, where the market features thousands of stocks hitting daily price-drop limits and experiencing trading halts. See, e.g., https://zhuanlan.zhihu.com/p/38345689.

upswing periods of a volatile market.[25] Finally, more historical media coverage induces a more negative tone; we conjecture that this is due to the media's tendency to follow up on negative news to a greater extent than on positive news. The drivers of *PoliticalPos* are, by and large, the same as *Pos*, which reflects the positive correlation (0.35) shown in Panel B of Table III between these two sentiment measures.

As might be expected, past returns of all horizons are significantly negatively related to *Neg_net*, and positively related to *Pos*; in other words, lower past long- and short-term returns both lead to more negative tone in the news. While it is unlikely that the current news reflects long-term past returns, it is likely that, over the short term, there is a leakage of news that is manifested in previous-day returns. In Section 3, we further examine the correlation of returns with news article sentiment.

2.6 Validation of news sentiment at the article level

In this section, we offer a validation of our news sentiment measure at the article level in a number of ways. Following YZZ, we manually read a subset of our news sample and classify each article as either positive or negative. We read 5,000 news articles (until we reach an equal number of positive and negative articles); we also ensure that the news articles that we read are evenly distributed across year, industry, and firm size.

We first compare our dictionary-sentiment measure of *Neg_net* of these 5,000 articles with the article-level human label. Panel A of Table VI shows that, in the human-labeled negative (positive) news, 85.88% (88.40%) of the articles have a *Neg_net* no-smaller (smaller) than zero. In aggregate, 87.14% of the news articles as judged by *Neg_net* are consistent with the article-level human classification. The 12.86% of articles that are misclassified by *Neg_net* are more difficult to interpret, in that they are more likely to feature more comparable levels of *Neg* and *Pos*. For instance, if we define a news article to be "complex" when both its *Neg* and *Pos* are either above or below the corresponding sample medians, 68% of the misclassified articles are complex, as compared to only 31% for the correctly identified articles.

[Insert Table VI about here.]

[25] For example, data from the Federal Reserve Bank of St. Louis show that, during our sample period, the CBOE China ETF Volatility Index (VIX) has a mean value of 24.46, as opposed to the mean US VIX of 14.89.

We next compare the dictionary-word counting *Neg_net* of these articles with the more traditional supervised machine learning method of support vector machine (SVM), a classifier that divides the sample into distinct groups by maximizing the distance (i.e., distinctiveness) among the groups (e.g., Cortes and Vapnik, 1995; Vapnik, 2013). In our setting, SVM classifies each article into either positive or negative news based on a training sample.

We divide the 5,000 articles into training and test sets; we use 70% of the articles as our training set, and the rest as the test set. First, we evaluate the accuracy of the SVM for the test sample. Panel B of Table VI reports that the overall accuracy of the SVM classification for the test set (the "Weighted F-1-score") is 88.20%—that is, using the SVM separation rule obtained in the training set, we are able to correctly identify the article sentiment in the test set about 90% of the time. This compares favorably with existing studies that typically have a 60-90% accuracy rate for SVM in Chinese texts. For example, Huang et al. (2014, Table 10) count the sentiment words and emojis of over two billion Weibo posts (similar to Twitter tweets) based mostly on the Hownet sentiment dictionary (a generic Chinese dictionary), and report SVM accuracy rates of 75-78% for negative/positive classification of the posts. Moreover, our SVM accuracy rate is highly similar to the 87.14% matching rate between *Neg_net* and human-labeling in Panel A. That is, our dictionary-based sentiment word counting achieves a highly similar rate of accuracy as traditional machine learning classifiers such as SVM. Finally, within the test set, we compare the SVM predicted article-level sentiment with that judged by *Neg_net*. We find that 83.60% of the test-set articles are consistently judged by SVM and *Neg_net*.

In Panel B, we also utilize, respectively, 60% and 80% of the 5,000 articles as the training set, and find that the results are highly similar. In sum, Table VI shows that our dictionary-based sentiment measure matches, in about 90% of cases, the article-level human interpretation, and achieves an accuracy rate comparable to the article-level machine learning classifier of SVM. Inherently, the dictionary method comes with the advantages of being simple and straightforward to implement, and offers a continuous sentiment evaluation (as compared to the largely binary evaluation offered by SVM).

Lastly, as another article-level validation, we utilize the Wind Terminal, which labels news articles to be either positive, negative, or null (which we interpret as neutral). In untabulated results, we download 50,000 articles from the Wind Terminal, and find that 86.75% of the articles that are labeled as positive or negative are consistently judged by *Neg_net*. Again, the misclassified articles tend to be more complex. While the Wind Terminal's article classification algorithm is unknown to us, we take comfort that our dictionary-based sentiment judgment achieves an identification level similar to our manual reading but on a much larger corpus size from Wind.

3. The Association of News Sentiment and Returns

Given the close relationship between our news sentiment measures and past returns, in this section, we carry out an in-depth analysis of the association between news sentiment and past and future stock returns. This analysis further helps to determine the efficacy of our new Chinese dictionary.

3.1 Return Implications

As previously discussed, we map the day [0] return to news released between the end of the previous trading day and the end of the current trading day. Therefore, if the news is released pre-market (during the middle of the trading day), the day [0] return reflects the immediate effect of the news for the full (partial) trading day. We note that 88.0% of news articles are released on trading days, with the rest being released during holidays or weekends. Among the news released during trading days, Figure 2 depicts the frequency distribution of the publication time stamps; we note that 16.4% of those are published during trading hours.[26] In the Internet Appendix, we offer a robustness check by removing all intra-trading-day news releases, so that day [0] contains the

[26] There is a large spike of news releases at 00:59 and 01:00; these consist mainly of scheduled firm-initiated announcements at regulatory-designated media outlets. This contrasts with the regularity found in U.S. news releases due to firm-initiated announcements, which exhibit spikes at the top of each hour and each half-hour during the trading day (HTW).

entire trading day for a given stock following a news release (instead of a partial trading day), and find that our conclusions remain the same.

[Insert Figure 2 about here.]

We regress historical and future returns from day [-10] to day [10] on our news sentiment measures. As with Table V, we control for the Chinese version of the Fama-French three factors and other past stock and news attributes, all measured at or before day [-11]. In particular, the control variables include the Chinese market's "CH-4" factors of market, size, value, and turnover in Liu, Stambaugh, and Yuan (2019), where the authors emphasize turnover as a measure for investor sentiment, as the market is dominated by individual investor trading. To control for industry and firm size effects on returns, we adjust returns by returns on an industry- and size-matched portfolio.[27]

[Insert Table VII about here.]

Table VII reports the regression results. We first discuss the results on the control variables. The coefficient estimates of the control variables show a number of well-known patterns in Chinese markets, in that size, turnover, and firm age since IPO all strongly and negatively predict returns. In addition, there is a strong short-term momentum and immediate weekly reversal of returns; for example, the excess return between days [-5] and [-3] positively predicts returns on days [-2, -1], but negatively predicts returns on days [0, 10]. There is also a longer-term return reversal effect: returns from months [-12, -2] negatively predicts returns from days [-10, 1]. Overall, these control variable patterns on returns are consistent with that documented in the literature. For example, Liu, Stambaugh, and Yuan (2019) summarize the Chinese anomalies literature (see their Appendix A.2), and replicate the existence of anomalies such as size, value, volatility, turnover, and reversal. The weekly return reversal effect is also documented in studies such as Chen, Hua, and Jiang (2018), Gang, Qian, and Xu (2019), and Yu, Fung, and Leung (2019).

[27] To construct the industry- and size-matched portfolios, we break each industry in Table IV into three size terciles, then use the value-weighted portfolio return for each tercile as the industry- and size-matched portfolio return for a stock that is a member of that portfolio. The results remain qualitatively similar if we use simple excess market returns.

Turning to sentiment, Table VII (Panels A and B) shows that our sentiment measures are associated with past and future returns. *Neg_net* and *Neg* are negatively, and *Pos* is positively associated with excess returns from days [-10] to [1]. The positive association between returns and *PoliticalPos* is only positive and significant over days [-2] to [1], and is, thus, much weaker. Moreover, these returns are insignificant on day [2], before they are largely reversed over days [3, 5] and [6, 10]. The reversal in returns that appears for all four of the sentiment-based measures is consistent with the control variable results in Table VII of return reversals.

One way to gauge the efficacy of our dictionary is to examine the economic significance of sentiment on returns. To this end, we note that the coefficient estimates of sentiment on returns are an order of magnitude larger on days [-1] and [0] than on other days, with the largest estimate on day [0]. For example, the *Neg_net* coefficient estimates on returns on days [-1] and [0] are, respectively, −6.787 and −11.366, as compared to, for instance, −2.171 on day [-2]. Using the standard deviation of the variable times the variable's fitted coefficient estimate as the measure of economic significance, *Neg_net*'s one standard deviation impact on returns is −0.51% (= −11.366% × 0.0450) on day [0] and −0.31% on day [-1]; this economic significance amounts to a highly significant level on an annualized basis.

Table VII suggests that the statistical significance of sentiment on returns may extend beyond the range of [-10, 10] days. In Figure 3, we plot the coefficient estimates of *Neg_net* in return regressions for each day during days [-20, 20], based on the specification of Table VII. We observe that the significantly negative association between *Neg_net* and return starts at day [-20], and the significantly positive association can last to day [10]. We note that the magnitudes of the coefficient estimates on days [-1] and [0] are considerably larger; untabulated, tests reject that the coefficient estimates on days [-1] and [0] are equal to those of prior days, and also reject that the coefficient estimates on days [2, 20] are equal to those on days [-20, -2]. Outside of days [-4, 1], the coefficient estimates (albeit many of them are still statistically significant) are close to zero. For example, on day [-5], the coefficient estimate is only about one-eighth of its value on day [0]; accordingly, *Neg_net* carries an economic significance of only −0.066% (−6.6 bps) on returns on that day.

[Insert Figure 3 about here.]

Is the statistical significance, outside of day [-1, 0], actually tradable? The right axis of Figure 3 plots the corresponding economic significance (the coefficient estimate multiplied by the standard deviation of *Neg_net*) of *Neg_net* for days [-20, 20]. We observe three tiers of economic significance: days [-1, 0] at the range of 30-50 bps/day, days [-4, -2] at around 10 bps, and the rest smaller than 7 bps—with many very close to zero. During our sample period, there are two major components of explicit trading costs in Chinese stock markets: a government-imposed 0.1% stamp tax for a round-trip trade, and a commission payable to the broker that averages around 0.05% in each direction of the trade.[28] These costs sum to 20 bps for each round-trip trade. It, thus, appears that only the first tier (i.e., days [-1,0]) among these three tiers of economic significance is tradable when considering a round-trip.

Analogously, we find that the economic significance of *Neg*, *Pos*, and *PoliticalPos* is, respectively, 30.6, 44.3, and 16.5 bps on day [0], and 11.7, 30.9, and 7.9 bps on day [-1]. A further comparison of the economic significance between the *Neg* and *Pos* sides reveals that the difference is significant at the 1% level for both days [0] and [-1]. The economic significance is not economically meaningful for all other days for these measures. In sum, the results in Table VII indicate that our news sentiment measures are associated with returns on days [-1, 0] with a strong economic magnitude, that *Pos* has a stronger effect than *Neg*, and that *PoliticalPos* has a much weaker return association (consistent with politically-inclined positive words being disregarded by most investors). The significance of Chinese *Pos* contrasts with the findings in the US market (in the English-language setting), where positive news tone is either insignificant or less significant than negative tone in predicting returns (e.g., Tetlock et al., 2008; HTW). Compared to US markets, the significance of positive news tone in China, thus, seems unique.[29]

3.2 Does there exist an information leakage of news?

[28] The average one-way commission is 4.94 bps from 2013-2019, based on Securities Association of China (2020).

[29] We believe that this is a new finding made possible by our new Chinese financial dictionary, and warrants further research on cultural differences in China vs. the U.S.

That *Neg_net* is significantly related to returns from days [-4, -1], with an economic significance of 10 bps and above, suggests that there may exist information leakage of news before it is released. In fact, information leakage is believed to be widespread in China and has been documented in various settings; for example, institutional investors trade ahead of announcements of the firm's split-share reform (a reform to float all classes of shares in 2005-2008) (Tong, Zhang, and Zhu, 2013), short-selling activities increase prior to poor earnings announcements (Feng and Chan, 2016), and super-rich investors trade profitably around dividend announcements of local firms (Li, et al., 2017).[30]

Yet, as discussed in HTW, it is important to control for potential persistence in news sentiment when examining the direction of causality between news releases and stock returns; for instance, news may be slowly released prior to a major news release at a particular date, and our news source may not completely capture such prior news releases. Accordingly, we investigate the non-information-leakage possibility by adjusting for the persistence of news sentiment. Untabulated, we can report that a firm's day [0] *Neg_net* has an autocorrelation coefficient of 0.15-0.20 with the firm's past half-year *Neg_net*; for example, day [0] *Neg_net* has an autocorrelation coefficient of 0.17 with *Neg_net* of days [-40, -21]. Although the autocorrelations are relatively small, they are statistically significant. We ameliorate the news persistence by normalizing the sentiment measures—in essence, measuring the unexpected "shock" in news. Specifically, for each sentiment measure, we compute a standardized abnormal measure by subtracting, at the firm level, the measure's past six-month mean, then divide the difference by the measure's standard deviation during the period (e.g., Tetlock et al., 2008). Table VIII regresses returns on these abnormal sentiment measures, where we continue to include all of the control variables shown in Panel A of Table VII. We observe that the abnormal *Neg_net* is now negatively and significantly related to returns only on days [-1], and [0], and the sign reverses for days [2, 10]. More conspicuously, the magnitudes of the coefficient estimates are much larger on days [-1] and [0] than on all other days, consistent with Table VII.

[30] See, for example, a 2019 white paper on information transparency in China released by Guanghua-Rotman Centre for Information Capital Market Research (https://www.rotman.utoronto.ca/-/media/orphan-files/China-White-Paper.pdf).

[Insert Table VIII about here.]

These results indicate that the adjustment to reduce the effect of news persistence does not rule out the economic significance of the effect of sentiment on returns on day [-1]. This finding suggests that there exists some degree of news leakage in China.[31] We, however, emphasize that our sentiment measures induce a significant economic impact on day [0], in that the return impact is much larger than any other day surrounding the news, including day [-1].

3.3 "Horse race" with alternative dictionaries in return associations

In this section, we compare the return implications of our dictionary with those of the other Chinese-language financial dictionaries. As previously discussed, we consider three alternative dictionaries: the LM dictionary, the YZZ dictionary, and the common set of three general dictionaries (the Generic dictionary). Earlier, in Table II, we showed that the sentiment measures across these four dictionaries are significantly correlated, and, hence, it is possible that the return associations with our dictionary's sentiment measures might be subsumed by these alternative dictionaries. Indeed, in the Internet Appendix, we report that, when each of these alternative dictionaries is used individually, the sentiment measures show a similar pattern of returns to that in Table VII; that is, the negative (positive) sentiment measures are generally statistically negatively (positively) related to returns for windows during days [-10, 1], and the return effect is largely reversed for windows during days [2, 10].

In Table IX, we carry out a horse race among the sentiment measures. In the table, we include *Neg_net*, *Neg_net_LM*, *Neg_net_YZZ*, and *Neg_net_generic* simultaneously in our return regressions that also include control variables, as described in Panel A of Table VII. We observe that, not only is the statistical significance of *Neg_net* preserved in return regressions for windows during days [-10, 10], but the signs of both *Neg_net_LM* and *Neg_net_generic* are reversed, compared to their standalone coefficient values. *Neg_net_YZZ* remains significantly negative for all windows between days [-10, 1]; however, the magnitude of its estimate is substantially smaller

[31] We note that this result is not likely to be caused by systematically erroneous time stamps in our news sample. Specifically, we compare our news time stamps with those in the standard commercial news feed of the Wind Terminal for all news for 10 randomly chosen firms. The comparison shows that our time stamps are identical to those in the Wind Terminal.

than that of its standalone regressions (reported in the Internet Appendix) and than that of *Neg_net*. In sharp contrast, the magnitudes of the *Neg_net* coefficient estimates are only slightly reduced from their standalone values (in Table VII). In sum, Table IX indicates that *Neg_net* subsumes all of the return relations of *Neg_net_LM* and *Neg_net_generic*, as well as much of that of *Neg_net_YZZ*.

[Insert Table IX about here.]

Given that our dictionary incorporates much of YZZ's vocabulary, it is perhaps not surprising that the return associations of the YZZ dictionary are much reduced. It is still possible that the return implications of our dictionary stem from the YZZ vocabulary in our dictionary. To explore this possibility, we re-construct the sentiment measures using the dictionary derived from Word2vec based on seed words from reading only the 2,500 news articles (Iterations 1-4 in Panel B of Table I); that is, in this dictionary, we rely only on seed words from our reading of 2,500 articles—but not seed words from YZZ. In Panel A of Table X, we run the return regressions with this *Neg_net* (dubbed *Neg_net_2500*) and *Neg_net_YZZ* side-by-side. Again, we find that *Neg_net_YZZ*'s magnitude of coefficient estimates remain small for all periods during the event window. In contrast, we continue to observe *Neg_net_2500* to exhibit negative and significant coefficient estimates on days [-1, 0], with much larger magnitudes compared to other days. Thus, the outperformance of our dictionary is also present for our vocabulary outside of YZZ.

[Insert Table X about here.]

Lastly, it is possible that our outperformance over the YZZ dictionary arises chiefly from either the common or non-common ordinal ranking of news articles. To reflect the overlap of ordinal rankings by these two dictionaries, we categorize articles into "agreeing" and "disagreeing" articles using a comparison between *Neg_net_2500* and *Neg_net_YZZ*. We define agreeing articles as those ranked by both dictionaries to be either above-median tone or below-median tone; the remainder are defined as disagreeing news. Given the strong correlation between the two dictionaries, we find that there is more agreeing news: 85.9% of the news are agreeing news.

Panel B of Table X runs the return regressions separately for both types of news, with sentiment measures from both dictionaries included as explanatory variables. For agreeing news, we find that *Neg_net_2500* subsumes *Neg_net_YZZ* on day [0]; further, examining both the negative and positive tone sides, while *Neg_YZZ* remains significant, the sign of *Pos_YZZ* reverses to become negative for days [0] and [1]—consistent with our prior finding that the YZZ dictionary contains many positive words that are political words. For disagreeing news, while *Neg_net_YZZ* retains significance on days [-1] and [0], its magnitude of coefficients is much smaller than that of *Neg_net_2500*. In contrast, for both agreeing and disagreeing news, not only does our dictionary retain its statistical significance, but its economic significance is also not much compromised when compared to the standalone case in Table VII. In sum, these results suggest that, while the YZZ sentiment retains some explanatory power for returns incremental to our dictionary, the economic significance of YZZ is weaker than ours. For the majority of news articles, where our and YZZ dictionaries agree on ordinal ranking, our dictionary subsumes YZZ in return associations, especially on the positive side of news sentiment. The associations between returns and our sentiment measures are, therefore, cardinal, and not simply ordinal.

4. Politically Inclined Words and Media Bias

One innovation in this paper is that we separate politically-inclined positive words from words that convey normal positive sentiment. Politically-inclined positive words, while still reflecting sentiment, manifest "Party-line" journalism (e.g., Qin**,** Strömberg, and Wu, 2018) — notably, propaganda—advocated and often mandated by the state. In fact, Qin**,** Strömberg, and Wu (2018) propose to measure media bias in China based on the coverage of "government mouthpiece" content such as the number of mentions of party leaders and the number of cites of Xinhua News Agency (the official central government news agency) within the news, relative to commercial content. In a similar vein, Piotroski, Wong, and Zhang (2017) find that articles published by commercially-oriented business newspapers are less politically biased than those published by official newspapers. Piotroski, Wong, and Zhang (2017) tag an article's political bias by the frequency of political phrases in the Dictionary of Scientific Development (Xi, 2007). This dictionary contains 1,682 words that are mostly proper nouns and, thus, have little sentiment connotation.[32] Our list of politically-inclined positive words not only reflects the media bias from party-line journalism, but also complements Piotroski, Wong, and Zhang (2017) with a list of political words that embody sentiment.

In addition to having more political words relative to U.S. news articles, media bias in China exhibits other patterns. Most notably, both Piotroski, Wong, and Zhang (2017) and YZZ report that state media outlets (relative to business or market media outlets) issue fewer negative corporate news articles. This arises either because stake-holding politicians desire to avoid political costs such as diminishing career prospects caused by negative news, or the party-line in general wishes to avoid public dissent and political de-stability in such news (e.g., Piotroski, Wong, and Zhang, 2017). These authors subsequently report that news stories by state media have lower value relevance, including a smaller price impact on corresponding stocks.

In light of the above literature, in this section, we examine the potential of using our political and negative words as a sentiment-related measure for media bias. We ask two questions: i) does state media use more politically-inclined positive words and fewer negative words, and ii) do sentiment measures from state media exhibit a lower association with returns?

To begin with, we classify our sample news outlets as state media vs. non-state media. Following YZZ, we define a media outlet to be state media if its ultimate control right belongs to the central or a provincial government. Unlike aggregating articles at the firm-day in our prior return analysis section, here we keep all firm-specific news articles as the tone of each news article may be tied to its authoring source. Retaining other news filtering procedures as in our prior Section 3, we have 682,945 news articles and 64 news outlets. Three-quarters of the news outlets are state media, consistent with the fact that most of the media outlets are controlled by either the central or a provincial government; and the state media, in total, authors 78.2% of the articles in our sample, consistent with the literature (e.g., YZZ).

Table XI confirms that state media uses more politically-inclined positive words and fewer negative words. In the Table, we regress *PoliticalPos* and *Neg* on the state media dummy (in addition to control variables). We find that the state media dummy is significantly positive

[32] For the list of Xi (2007) words, see, e.g., https://xuewen.cnki.net/detail-r201509408.html.

(negative) on *PoliticalPos* (*Neg*). In both cases, state media's sentiment is about 10% different from non-state media; for example, state media has a *PoliticalPos* value of 0.258% higher than non-state media, against an overall sample mean for *PoliticalPos* of 3.0% (Table III). Given the literature's emphasis on state media being more political and less negative, we aggregate the two effects to create a media bias index by *PoliticalPos* minus *Neg*. Table XI shows that the coefficient estimate of state media on this media bias index, *MediabiasIndex*, equals the sum of the magnitude of the coefficient estimates on *PoliticalPos* and *Neg*, suggesting that state media's sentiment biases in *PoliticalPos* and *Neg* are largely distinct from each other.

[Insert Table XI about here.]

To provide a benchmark for the sentiment bias by state media, we follow Piotroski, Wong, and Zhang (2017) to measure political bias as the fraction of political phrases in the Dictionary of Scientific Development (Xi, 2007) for each article, dubbed *PoliticalNouns*. As previously mentioned, these political phrases are mostly proper nouns. They rarely appear in our dictionary: out of the 1,682 phrases in Xi (2007), only 52 (30) phrases appear in our sentiment word (political word) list. The last column of Table XI shows that state media uses 0.615% more *PoliticalNouns*; which is also about 10% of *PoliticalNouns*'s sample mean of 5.17%. Therefore, state media's sentiment bias is comparable to its bias in using political nouns.

In Table XII, we examine whether state media's sentiment is less associated with returns. Given our previous evidence that the sentiment effect on returns concentrates on days [-1] to [1] (and to a lesser degree days [-2, 2]), in this table, we regress returns over this period on the interaction term of sentiment and state media. Following the literature, we expect this interaction term to be significant, but in the opposite direction to the main sentiment effect, indicating a smaller association of returns with state media sentiment. Table XII confirms that state media sentiment is indeed less associated with returns. For instance, the interaction term of state media and *PoliticalPos* is negatively (and significantly) related to the cumulative return over days [-1, 1], while the main sentiment effect of *PoliticalPos* is significantly positive. For *PoliticalPos*, *Neg*, and *MediabiasIndex*, the magnitude of the coefficient estimate of the interaction term is about half that of the coefficient estimate of the sentiment by itself. In other words, the results suggest that the association between state media's sentiment and returns is about half of that of non-state media's. Therefore, stock prices respond less to sentiment words from state media as they do to non-state media. Again, to provide a benchmark, we examine the return associations due to state media's political bias *PoliticalNouns*. Table XII shows that state media's *PoliticalNouns* is less related to returns than that of non-state media, consistent with the literature (e.g., YZZ; Piotroski, Wong, and Zhang, 2017). Moreover, the magnitude of the coefficient estimate of the interaction term between state media and *PoliticalNouns* is even larger than that of *PoliticalNouns* itself. Aggregating these two coefficients would be the overall effect of state media's *PoliticalNouns* effect—which suggests that state media's *PoliticalNouns* effect is reversed to become negative, contrary to the desired effect of using political terms—an interesting endogeneity that deserves further research.

[Insert Table XII about here.]

In the remainder of Table XII, we place the interaction term between state media and sentiment and the interaction term between state media and *PoliticalNouns* in return regressions. We note that the significance of the state media and *PoliticalNouns* interaction is either subsumed (e.g., by the state media and *PoliticalPos* interaction), or reversed (e.g., by the state media and *Neg* interaction); in contrast, the interaction terms between state media and sentiment retain their magnitudes of estimates, for sentiment measures of *PoliticalPos*, *Neg*, and the media bias index. These results hold for cumulative returns over both [-1, 1] and [-2, 2]. In sum, Table XII shows that returns are less sensitive to sentiment bias in state media, but also shows that state media's sentiment bias is distinct from its bias in mentions of political entities or nouns.

We also single out a number of media outlets that are directly operated by the central government, such as Xinhua News Agency, People's Daily, Economic Daily, and China Central Television. Untabulated, we repeat the exercises in Tables XI and XII for these media outlets. As expected, they are more exuberant than other state media by exhibiting a higher value in *PoliticalPos*; in addition, these media outlets' political sentiment is less related to returns than other outlets.

5. Conclusion

In this paper we construct a Chinese financial sentiment dictionary from a large sample of corporate financial news using supervised machine learning of word embeddings. We generate a starting list of sentiment words by manually reading a subset of randomly chosen articles. We then iteratively train a neural network model to find semantically similar words. Reflecting the domain-specificity of finance, our process features representative seed lists of sentiment words by reading 2,500 news articles and the inclusion of the sentiment dictionary from You, Zhang, and Zhang (2018, "YZZ"), and also human-expert reviews of each iteration that in aggregate filter out 80% of the output words. We compare our dictionary with existing dictionaries that are largely based on the translation of Loughran and McDonald (2011, "LM") English sentiment dictionary. We use three benchmark dictionaries: our translation of LM augmented by appropriate synonyms, YZZ's translation of LM augmented by a manual list of keywords, and the intersection of three general-purpose, generic, sentiment dictionaries adopted in Huang, Wu, and Yu (2019). Our dictionary has little overlap with these benchmark dictionaries with only about 40% of our sentiment words covered in these dictionaries combined.

In validity checks, news sentiment metrics measured from our dictionary conform to ebbs and flows of the market, and to the fact that riskier firms and firms under distress tend to have negative-toned news. When cross-checked on the article level, our dictionary scores the sentiment of articles consistently with i) human reading of full articles, ii) traditional machine learning method of support vector machine, and iii) article sentiment tag produced by the Wind Terminal (the most widely used professional financial information terminal in China).

We further validate our dictionary by examining associations between news sentiment and returns. We find that news sentiment predicts returns, with strong economic significance, on the news announcement day. Distinctive in the Chinese market, we document that news sentiment is related to returns with meaningful economic significance the day before the announcement, suggesting a limited degree of information leakage in China, and that positive news tone is no less important in its associations with returns than negative news tone. Moreover, our dictionary outperforms and subsumes the benchmark dictionaries in return associations. When we place our dictionary side by side with the benchmark dictionaries in return regressions, not only is the significance of our sentiment measures largely intact, but our sentiment measures also subsume or much reduce the significance of other dictionaries.

We also create a list of politically-inclined positive words to reflect the contemporaneous mainland culture of state-controlled media that emphasizes overall coherence with central government guidelines. The political words are positive linguistically and appear frequently in our news article sample. We show that political words however have non-unique return associations and merit standalone consideration than general positive words. We demonstrate that state media exhibits a sentiment bias by using more politically-related positive and fewer negative words; and this bias renders state media's sentiment less informative on stock returns. State media's sentiment bias is distinct from its bias in the mentions of political entities or nouns as previously documented in the literature (e.g., Piotroski, Wong and Zhang, 2017), highlighting the necessity to isolate the list of political words when studying the news sentiment in China.

Our study highlights the importance of language and domain specificity in dictionary-based sentiment analysis. We show that direct translations of English list of sentiment words and a general list of Chinese sentiment words both have poor return associations in China. The emphasis on language specificity reflects that the fields of economics and finance are related to culture (e.g., La Porta et al., 1998) and that Chinese is distinct from English linguistically and culturally. The emphasis on domain specificity, on the other hand, roots from Loughran and McDonald's (2011) departure from a more general Harvard Psychosociological Dictionary in their creation of the finance-sentiment word list. With machine learning techniques being increasingly adopted in finance research, our paper joins the financial text processing literature showcasing the word embedding approach (e.g., Hanley and Hoberg, 2019; Cong, Liang, and Zhang, 2020); in particular, we utilize the finance domain specificity by human expertise reviews in the creation of our dictionary (see also Li, Mai, Shen, and Yan, 2021). Compared with other machine learning methods, sentiment judging from our dictionary is straightforward to implement, and is potentially readily applicable to other texts. As China arises to become one of the largest stock markets in the world, this paper's attempt to construct a context specific financial sentiment dictionary, we hope, carries economic importance in the field.

References

Bradshaw, M.T., Huang, A.G. and Tan, H. (2019) The effects of analyst-country institutions on biased research: Evidence from target prices, Journal of Accounting Research 57, 85–120.

Chen, Q., Hua, X. and Jiang, Y. (2018) Contrarian strategy and herding behaviour in the Chinese stock market, European Journal of Finance 24, 1552–1568.

Cong, L. W., Liang, T. and Zhang, X. (2020), Textual factors: A scalable, interpretable, and data-driven approach to analyzing unstructured information, unpublished working paper, Cornell University, University of Chicago, University of Chicago.

Cortes, C. and Vapnik, V. (1995) Support-vector networks, Machine Learning 20, 273-297.

De Jong, E. (2009) Culture and Economics: On Values, Economics and International Business, Routledge, London and New York.

Fama, E.F. and French, K.R. (1993) Common risk factors in the returns on stocks and bonds, Journal of Financial Economics 33, 3–56.

Fedyk, A. and Hodson, J. (2020), When Can the Market Identify Old News? unpublished working paper, Harvard Business School, Knowledge Research Group, Bloomberg L.P.

Feng, X. and Chan, K.C. (2016) Information advantage, short sales, and stock returns: Evidence from short selling reform in China, Economic Modelling 59, 131–142.

Gang, J., Qian, Z. and Xu, T. (2019) Investment horizons, cash flow news, and the profitability of momentum and reversal strategies in the Chinese stock market, Economic Modelling 83, 364–371.

Hanley, K. W. and Hoberg, G. (2019) Dynamic interpretation of emerging risks in the financial sector, The Review of Financial Studies 32, 4543-4603.

Huang, A.G., Tan, H. and Wermers, R. (2020) Institutional trading around corporate news: Evidence from textual analysis, The Review of Financial Studies 33, 4627–4675.

Huang, A.G., Wu, W. and Yu, T. (2019) Textual analysis for China's financial markets: A review and discussion, China Finance Review International 10, 1–15.

Huang, M., Ye, B., Wang, Y., Chen, H., Cheng, J. and Zhu, X. (2014) New word detection for sentiment analysis, In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 531-541.

Jurafsky, D. and Martin, J.H. (2019) Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (third edition draft).

Kothari, S.P., Shu, S. and Wysocki, P.D. (2009) Do managers withhold bad news? Journal of Accounting Research 47, 241–276.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A. and Vishny, R.W. (1998) Law and finance, Journal of Political Economy 106, 1113–1155.

Li, J., Chen, Y., Shen, Y., Huang, Z. and Wang, J. (2019), Measuring China's stock market sentiment, unpublished working paper, Duke University, National Peking University, National Peking University, National Peking University, National Peking University.

Li, K., Mai, F., Shen, R. and Yan, X. (2021) Measuring corporate culture using machine learning, The Review of Financial Studies 34, 3265-3315.

Li, X., Geng, Z., Subrahmanyam, A. and Yu, H. (2017) Do wealthy investors have an informational advantage? Evidence based on account classifications of individual investors, Journal of Empirical Finance 44, 1–18.

Lin, L. and Xie, D. (2016) Do investors listen for the meanings behind executives' words? An empirical analysis based on management tones (translated from mandarin), Journal of Finance and Economics 7, 28-39.

Liu, J., Stambaugh, R.F. and Yuan, Y. (2019) Size and value in China, Journal of Financial Economics 134, 48–69.

Loughran, T. and Mcdonald, B. (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, Journal of Finance 66, 35–65.

Mao, H., Gao, P., Wang, Y. and Bollen, J. (2014) Automatic construction of financial semantic orientation lexicon from large-scale Chinese news corpus, Institut Louis Bachelier 20, 1-18.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) Efficient Estimation of Word Representations in Vector Space. 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings, 1-12.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013) Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems, 3111–3119.

Piotroski, J. D., Wong, T. J. and Zhang, T. (2017) Political bias in corporate news: The role of conglomeration reform in China, The Journal of Law and Economics 60, 173-207.

Qin, B., Strömberg, D. and Wu, Y. (2018) Media bias in China, American Economic Review 108, 2442-2476.

Robinson, R.C. (2018) A linguistics approach to solving financial services standardization, Journal of Financial Market Infrastructures 7, 61–72.

Stulz, R.M. and Williamson, R. (2003) Culture, openness, and finance, Journal of Financial Economics 70, 313–349.

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

Tetlock, P.C. (2011) All the news that's fit to reprint: Do investors react to stale information, Review of Financial Studies 24, 1481–1512.

Tetlock, P.C., Saar-Tsechansky, M. and Macskassy, S. (2008) More than words: Quantifying language to measure firms' fundamentals, Journal of Finance 63, 1437–1467.

Tong, W.H.S., Zhang, S. and Zhu, Y. (2013) Trading on inside information: Evidence from the share-structure reform in China, Journal of Banking and Finance 37, 1422–1436.

Vapnik, V. (2013) The Nature of Statistical Learning Theory, Springer Science & Business Media, Berlin.

Wang, C. and Wu, J. (2015) Media tone, investor sentiment and IPO pricing (translated from mandarin), Journal of Financial Research 9, 174-189.

Xi, J. (2007) Dictionary of Scientific Outlook on Development, Shanghai Lexicographical Publishing House, Shanghai.

Xie, D. and Lin, L. (2015) Do management tones help to forecast firms' future performance: A textual analysis based on annual earnings communication conferences of listed companies in China (translated from mandarin), Accounting Research 2, 20-27.

Yan, Y., Xiong, X., Meng, J.G. and Zou, G. (2019) Uncertainty and IPO initial returns: Evidence from the Tone Analysis of China's IPO Prospectuses, Pacific-Basin Finance Journal 57, 101075.

Yao, J., Feng, X., Wang, Z., Ji, R. and Zhang, W. (2021) Tone, sentiment and textual analysis: The construction of Chinese sentiment dictionary in finance (translated from mandarin), Journal of Management Sciences in China 24, 26-46.

You, J., Zhang, B. and Zhang, L. (2018) Who captures the power of the pen? The Review of Financial Studies 31, 43–96.

Yu, L., Fung, H.G. and Leung, W.K. (2019) Momentum or contrarian trading strategy: Which one works better in the Chinese stock market, International Review of Economics & Finance 62, 87–105.

Zeng, Q., Zhou, B. and Zhang, C. (2018) Annual report tone and insider trading: Do insiders act as what they said (translated from mandarin), Management World 9, 143-160.

Zhong, K., Dong, X. and Chen, Z. (2020) The tone of the earnings communication conference and analyst forecast accuracy (translated from mandarin), Business and Management Journal 8, 120-137.

Zhou, B., Zhang, C. and Zeng, Q. (2019) Annual report's tone and stock crash risk: Evidence from China A-share companies (translated from mandarin), Accounting Research 11, 41-48.

Appendix A News Filtering and Sample Selection

We retrieve corporate news, dated as far back as possible, for all Chinese firms from finance.sina.com.cn as of August 2019. For the dictionary construction, we keep all of the news. For firm-specific news, we use the following data filtering process. In Step (2), the keywords include "Daily Collection of Firm Announcements", "Collection of Investment Opportunities", "Industry Research", "Collection of Stocks Hitting Daily Limits Today", "Stocks Sought After by Funds", "Highlights of Important Market News", "Industry Information Revealed", and "League Table of Stock Market." In Step (3), firm identity can be stock ticker code, firm full name or name abbreviation, and firm's URL, and we require that i) the firm tagged by finance.sina.com.cn is among the top three-mentioned firms, and ii), if the firm assigned is not the most mentioned, the frequency of its mentions is at least 50% of that of the top mention. In Step (5), we examine the headlines of the news articles in the surrounding [-3, 3] trading days. If the headlines of the two news articles have an overlap of at least 70%, we treat the later article as a reprint of the earlier article.

		# of news articles or stock-days left	# of firms
(1)	"Firm News" stories retrieved from finance.sina.com.cn, dated January 2013 to August 2019	3,078,175	3,557
(2)	Remove news whose headline contains keywords related to market or industry commentaries	2,594,817	3,557
(3)	Assign news to a given firm by counting the frequencies of firm-identify mentions	1,1135,28	3,553
(4)	Remove news articles with less than 50 words, excluding stop words (e.g., Tetlock et al., 2008)	967,919	3,552
(5)	Remove news that is reprint of an earlier news article	846,218	3,552
(6)	Combine news released on the same trading day for a given firm	499,903	3,552
(7)	Remove the first 20 days post the IPO date, stock-days designated as ST (special treatment), trading halts, and the first day after more than one day of trading halt	424,758	3,539

Appendix B Variable Definitions

Variable	Definition
Neg_net; *Neg_net_2500*	The fraction of total negative word count net of total positive word count relative to the total number of words in a news article, based on our full dictionary (*Neg_net*) or the dictionary constructed from reading 2,500 articles (*Neg_net_2500*).
Neg; Neg_2500	The fraction of total negative word count relative to the total number of words in a news article, based on our full dictionary (*Neg*) or the dictionary constructed from reading 2,500 articles (*Neg_2500*).
Pos; Pos_2500	The fraction of total positive word count relative to the total number of words in a news article, based on our full dictionary (*Pos*) or the dictionary constructed from reading 2,500 articles (*Pos_2500*).
PoliticalPos; *PoliticalPos_2500*	The fraction of total politically-inclined word count relative to the total number of words in a news article, based on our full dictionary (*PoliticalPos*) or the dictionary constructed from reading 2,500 articles (*PoliticalPos_2500*).
Ab_Neg_net	*Neg_net* subtracted by its past six-month mean (excluding the first 20 days prior to the news), then divided by its standard deviation during the period. Missing mean and standard deviation are replaced by those estimated in months −12 to −7. *Ab_Neg*, *Ab_Pos*, and *Ab_PoliticalPos* are similarly defined on, respectively, *Neg*, *Pos*, and *PoliticalPos*.
Neg_net_LM	The fraction of total negative word count net of total positive word count relative to the total number of words in a news article, based on an expanded version of the Chinese translation of the Loughran and McDonald (2011) dictionary.
Neg_LM	The fraction of total negative word count relative to the total number of words in a news article, based on an expanded version of the Chinese translation of the Loughran and McDonald (2011) dictionary.
Pos_LM	The fraction of total positive word count relative to the total number of words in a news article, based on an expanded version of the Chinese translation of the Loughran and McDonald (2011) dictionary.
Neg_net_YZZ	The fraction of total negative word count net of total positive word count relative to the total number of words in a news article, based on the You, Zhang, and Zhang (2018) dictionary.
Neg_YZZ	The fraction of total negative word count relative to the total number of words in a news article, based on the You, Zhang, and Zhang (2018) dictionary.
Pos_YZZ	The fraction of total positive word count relative to the total number of words in a news article, based on the You, Zhang, and Zhang (2018) dictionary.
Neg_net_generic	The fraction of total negative word count net of total positive word count relative to the total number of words in a news article, based on the common set of sentiment dictionaries by Dalian University of Technology, HowNet, and National Taiwan University (the "Generic dictionaries").

Neg_generic	The fraction of total negative word count relative to the total number of words in a news article based on the Generic dictionaries.
Pos_generic	The fraction of total positive word count relative to the total number of words in a news article based on the Generic dictionaries.
PoliticalNouns	The fraction of total count of political phrases in the Dictionary of Scientific Development (Xi, 2007) relative to the total number of words in a news article.
beta	The firm's most recent CAPM beta measured using the monthly returns over months [-37, -2].
Log market cap.	The logarithm of market capitalization at the end of the quarter prior to the news.
Book to market	Book to market ratio at the end of the quarter prior to the news release date.
Turnover	The average daily stock turnover ratio (overall CSMAR market trading volume/shares outstanding) over the trading days of -32 to -11 relative to the news event day.
Volatility	The standard deviation of stock returns over days -32 to -11, relative to the news event day.
SUE	Standardized unexpected earnings of the most recent quarter, defined as the difference of earnings between the quarter and the quarter of a year ago, scaled by standard deviation of earnings of the previous 12 quarters.
Dividend yield	The trailing 12-month dividend yield of the previous month (past 12-month dividend divided by the beginning-of-the-month market capitalization).
Stock age	The logarithm of the number of months of the firm since IPO.
CSI300 dummy	A dummy variable that equals one if the stock is included in the CSI 300 index when the news is published.
SOE dummy	A dummy variable that equals one if the firm is a State-Owned-Enterprise in the prior calendar year.
Historical articles	The logarithm of one plus the number of articles mentioning the firm in the prior365 days. The value of year 2014 is used for the value of the first sample year 2013.
Number of articles$_t$	The logarithm of one plus the number of news articles in the day t.
Excess Return$_{t-1}$	Industry- and size-adjusted return of the stock at day $t-1$ (i.e., event day [-1]). This variable is similarly defined, by taking the mean daily value, over the horizons of past days $t-2$, $t-5$ to $t-3$, and $t-10$ to $t-6$; the current day (day [0]); and future days $t+1$, $t+2$, $t+3$ to $t+5$, and $t+6$ to $t+10$. Excess returns from months -12 to -2 ($m-12$ to $m-2$) are instead cumulative industry- and size-adjusted returns over the horizon. All returns are multiplied by 100.
State media	A dummy variable that equals one if the article is from a media outlet whose ultimate control right belongs to the central or a provincial government.
MediabiasIndex	*PoliticalPos* minus *Neg*.

Table I Construction of Sentiment Words Dictionary

This table shows the output during our construction of the sentiment words dictionary. In Panel A, we manually read 2,500 articles and also use the words in You, Zhang, and Zhang (2018) ("YZZ") to arrive at a list of starting words, which is in turn used in Word2vec in larger sets of articles to find synonyms (Panel B). In Panel B, "Seed articles" of 500 refers to articles used in Rounds 1-3 in Panel A, and "Seed articles" of 2,000 refers to articles used in Round 4 in Panel A. We manually screen the Word2vec synonyms to arrive at the final "valid synonyms." The number of unique words does not include stop words.

Panel A: Manually reading 2,500 articles for four rounds and utilizing the YZZ dictionary

			Sentiment words selected incrementally		
Round	# of news articles/source	# of unique words	Negative	Positive	Political
1	50	1,003	41	56	24
2	250	2,980	193	152	65
3	200	1,343	100	41	33
4	2,000	28,245	372	451	184
5	YZZ	--	264	133	16
Total	2,500+YZZ	33,571	970	833	322

Panel B: Synonyms produced by Word2vec and human review

				Additional synonyms from Word2vec			Additional valid synonyms		
Iter-ation	Seed articles/ words	# of firms	# of news articles	Negative	Positive	Political	Negative	Positive	Political
1	500	100	576,153	1,858	1,730	878	594	506	337
2	500	1,000	1,777,178	1,907	1,404	936	579	351	347
3	500	3,557	3,078,175	3,640	3,403	2,563	573	372	319
4	2,000	3,557	3,078,175	782	1,308	735	201	138	108
5	YZZ	3,557	3,078,175	648	1,077	148	69	35	6

Table II Sentiment Words in Our Dictionary versus Other Dictionaries

Panel A reports the total number of words in our sentiment dictionary. Panel B reports words from the Loughran and McDonald (2011) translation augmented by the Python Synonym package ("LM Translation"), the dictionary in You, Zhang, and Zhang (2018) ("YZZ Dictionary"), and the common set of sentiment words in three general dictionaries of Dalian University of Technology, HowNet, and National Taiwan University ("Generic Chinese Dictionary"). Panel C shows the number and percentage of our words that appear in these other dictionaries.

Panel A: Total number of words in our dictionary

	Negative	Positive	Political	Total
Total	2,986	2,235	1,439	6,660
Single Character	70	12	3	85
Two-Character	1,859	1,186	544	3,589
Three- and Four-Character	1,039	1,016	877	2,932
Five-Character and more	18	21	15	54

Panel B: Words in other dictionaries

	Negative	Positive	Total
LM Translation	1,337	327	1,664
YZZ Dictionary	1,583	1,425	3,003
Generic Chinese Dictionary	566	639	1,205
Total	3,034	2,108	5,130

Panel C: Overlapping of our dictionary with other dictionaries (percentage in parentheses)

	Negative	Positive	Political	Total
LM Translation	489 (16.38%)	144 (6.44%)	43 (2.99%)	676 (10.15%)
YZZ Dictionary	1,145 (38.35%)	812 (36.33%)	208 (14.45%)	2,165 (32.51%)
Generic Chinese Dictionary	134 (4.49%)	153 (6.85%)	74 (5.14%)	361 (5.42%)
Total	1,434 (48.02%)	910 (40.72%)	280 (19.46%)	2,624 (39.40%)

Table III The summary statistics and correlations of the sentiment measures

This table reports the summary statistics and the correlation matrix of the sentiment measures. See Appendix B for variable definitions.

Panel A: The summary statistics of the sentiment measures

	N	Mean	Std Dev	Minimum	Median	Maximum
Neg_net	424,758	-0.0216	0.0450	-0.1481	-0.0157	0.0941
Neg	424,758	0.0276	0.0237	0.0000	0.0223	0.1189
Pos	424,758	0.0492	0.0348	0.0000	0.0412	0.1600
PoliticalPos	424,758	0.0300	0.0242	0.0000	0.0241	0.1280
Neg_net_LM	424,758	0.0052	0.0274	-0.0594	0.0018	0.0925
Neg_LM	424,758	0.0272	0.0220	0.0000	0.0223	0.1053
Pos_LM	424,758	0.0219	0.0154	0.0000	0.0190	0.0741
Neg_net_YZZ	424,758	-0.0210	0.0489	-0.1538	-0.0152	0.1064
Neg_YZZ	424,758	0.0276	0.0251	0.0000	0.0206	0.1222
Pos_YZZ	424,758	0.0486	0.0369	0.0000	0.0393	0.1646
Neg_net_generic	424,758	-0.0054	0.0077	-0.0361	-0.0040	0.0114
Neg_generic	424,758	0.0024	0.0034	0.0000	0.0009	0.0159
Pos_generic	424,758	0.0077	0.0084	0.0000	0.0056	0.0410

Panel B: The correlation matrix of the sentiment measures

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1)	*Neg_net*	1.00												
(2)	*Neg*	0.65	1.00											
(3)	*Pos*	-0.86	-0.16	1.00										
(4)	*PoliticalPos*	-0.38	-0.22	0.35	1.00									
(5)	*Neg_net_LM*	0.60	0.45	-0.46	-0.38	1.00								
(6)	*Neg_LM*	0.52	0.53	-0.31	-0.20	0.83	1.00							
(7)	*Pos_LM*	-0.33	-0.06	0.39	0.40	-0.60	-0.06	1.00						
(8)	*Neg_net_YZZ*	0.83	0.53	-0.72	-0.56	0.65	0.51	-0.42	1.00					
(9)	*Neg_YZZ*	0.55	0.73	-0.21	-0.23	0.46	0.53	-0.08	0.68	1.00				
(10)	*Pos_YZZ*	-0.74	-0.21	0.81	0.58	-0.54	-0.32	0.51	-0.87	-0.22	1.00			
(11)	*Neg_net_generic*	0.20	0.07	-0.21	-0.23	0.12	0.06	-0.11	0.26	0.10	-0.28	1.00		
(12)	*Neg_generic*	0.41	0.30	-0.32	-0.01	0.31	0.39	0.00	0.37	0.27	-0.31	0.03	1.00	
(13)	*Pos_generic*	-0.02	0.06	0.07	0.21	0.02	0.09	0.10	-0.09	0.01	0.13	-0.92	0.37	1.00

Table IV Sentiment by Year and Industry

This table shows the mean values of our sentiment measures by year (Panel A) and industry (Panel B). We use the industry classification by the China Securities Regulatory Commission (CSRC). CSRC's Manufacturing industry accounts for about half of our sample size; we break CSRC's Manufacturing industry into the following six industries based on a mapping with the Global Industry Classification Standard (GICS): Capital Goods; Automobiles & Components; Consumer Durables & Apparel; Food, Beverage & Tobacco; Pharmaceuticals, Biotechnology & Life Sciences; and Technology Hardware & Equipment.

Panel A: Sentiment by year

Year	N	*Neg_net*	*Neg*	*Pos*	*PoliticalPos*
2013	7,637	-0.0247	0.0247	0.0494	0.0280
2014	82,339	-0.0241	0.0251	0.0491	0.0300
2015	84,067	-0.0226	0.0264	0.0490	0.0312
2016	65,967	-0.0179	0.0295	0.0474	0.0285
2017	55,431	-0.0253	0.0281	0.0534	0.0294
2018	53,468	-0.0227	0.0286	0.0512	0.0292
2019	75,849	-0.0173	0.0291	0.0464	0.0309
Total	424,758	-0.0216	0.0276	0.0492	0.0300

Panel B: Sentiment by industry

Industry	N	*Neg_net*	*Neg*	*Pos*	*PoliticalPos*
Consumer Durables & Apparel	15,756	-0.0161	0.0293	0.0454	0.0278
Pharmaceuticals, Biotechnology & Life Sciences	25,661	-0.0173	0.0278	0.0451	0.0293
Materials	64,600	-0.0173	0.0298	0.0470	0.0275
Mining	11,989	-0.0178	0.0325	0.0502	0.0279
Wholesale & Retail	20,845	-0.0190	0.0267	0.0457	0.0283
Real estate	24,117	-0.0199	0.0263	0.0462	0.0264
Agriculture	5,421	-0.0199	0.0317	0.0517	0.0285
Scientific Research & Technical Services	2,774	-0.0206	0.0265	0.0472	0.0316
Public Utilities	12,205	-0.0209	0.0255	0.0464	0.0301
Capital Goods	56,877	-0.0228	0.0263	0.0491	0.0318
Diversified	3,145	-0.0228	0.0250	0.0478	0.0291
Automobiles & Components	15,168	-0.0231	0.0290	0.0520	0.0330
Financial	32,361	-0.0231	0.0292	0.0523	0.0307
Technology Hardware & Equipment	37,659	-0.0239	0.0262	0.0501	0.0317
Construction	12,081	-0.0243	0.0232	0.0475	0.0307
Culture, Sports & Entertainment	7,349	-0.0244	0.0261	0.0506	0.0284
Transportation & Postal Services	13,853	-0.0244	0.0241	0.0485	0.0320
Food, Beverage & Tobacco	20,036	-0.0252	0.0297	0.0549	0.0292
Accommodation & Catering	3,129	-0.0254	0.0277	0.0531	0.0287
Leasing & Commercial Services	6,403	-0.0265	0.0271	0.0536	0.0323
Environment & Public Facility Management	5,063	-0.0271	0.0266	0.0537	0.0312
Software and Information Services	26,966	-0.0279	0.0261	0.0541	0.0332
Other	1,300	-0.0233	0.0261	0.0496	0.0337

Table V The Determinants of News Sentiment

The dependent variables are the sentiment measures, each multiplied by 100. All regressions include firm and individual month fixed effects. *T*-statistics are two-way cluster-adjusted at the firm and news-date levels, and are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	(1)	(2)	(3)	(4)
	Neg_net	*Neg*	*Pos*	*PoliticalPos*
beta	0.300***	0.202***	-0.098***	-0.081***
	(6.37)	(7.58)	(-2.89)	(-3.85)
Log market cap.	-0.297***	-0.050*	0.250***	0.084***
	(-5.55)	(-1.76)	(6.52)	(3.36)
Book to market	1.109***	0.616***	-0.496***	-0.266***
	(7.67)	(7.91)	(-5.00)	(-4.32)
Turnover	-3.946***	-0.299	3.681***	0.276
	(-4.78)	(-0.70)	(6.05)	(0.66)
Volatility	-6.985***	4.860***	11.869***	-2.570***
	(-4.72)	(5.86)	(10.09)	(-2.96)
SUE	-0.132***	-0.066***	0.066***	0.045***
	(-11.17)	(-10.02)	(8.26)	(8.65)
Dividend yield	-3.615**	-3.114***	0.543	0.000
	(-1.96)	(-3.38)	(0.39)	(0.00)
Stock age	-0.088	0.150***	0.235***	-0.104***
	(-1.24)	(3.45)	(4.54)	(-2.94)
CSI300 dummy	0.069	0.012	-0.060	-0.041
	(0.81)	(0.25)	(-0.95)	(-1.00)
SOE dummy	0.191	0.024	-0.165	-0.043
	(1.29)	(0.25)	(-1.61)	(-0.50)
Historical articles	0.301***	0.178***	-0.125***	-0.070***
	(8.83)	(9.24)	(-5.17)	(-4.41)
Number of articles$_t$	-0.186***	-0.104***	0.080***	0.022
	(-4.95)	(-5.42)	(3.05)	(1.08)
Excess Return$_{t-1}$	-0.157***	-0.034***	0.122***	0.023***
	(-34.07)	(-15.27)	(36.82)	(13.81)
Excess Return$_{t-2}$	-0.037***	0.002	0.038***	0.003*
	(-10.06)	(0.79)	(14.55)	(1.95)
Excess Return$_{t-5,t-3}$	-0.105***	-0.008**	0.097***	-0.000
	(-16.50)	(-2.49)	(19.67)	(-0.10)
Excess Return$_{t-10,t-6}$	-0.138***	-0.023***	0.116***	0.002
	(-17.66)	(-5.38)	(19.25)	(0.48)
Excess Return$_{m-12,m-2}$	-0.003***	-0.001***	0.002***	0.001***
	(-11.29)	(-6.40)	(11.26)	(4.89)
Constant	3.573***	1.920***	-1.714*	2.092***
	(2.74)	(2.80)	(-1.83)	(3.46)
Observations	410,943	410,943	410,943	410,943
Adj R-squared	0.142	0.095	0.140	0.097

Table VI Article-Level Validation of Word-Counting Sentiment

In Panel A, we manually read 5,000 articles that we label to be either negative or positive news, and match them against the article's *Neg_net* value. In Panel B, we employ support vector machine (SVM) for these 5,000 articles to classify an article in the test set to be either positive or negative. Precision is the ratio of true positives to the sum of true positives and false positives, recall is the ratio of true positives to the sum of true positives and false negatives, and F1-score is the harmonic mean of precision and recall.

Panel A: Article-level sentiment judgment by sentiment-word counting vs. by human

	# of articles	# of artciles by *Neg_net* value	Accuracy
Human-labeled negative news	2,500	2,147 with *Neg_net* ≥ 0	85.88%
Human-labeled positive news	2,500	2,210 with *Neg_net* < 0	88.40%
Overall	5,000	4,357	87.14%

Panel B: Article-level sentiment judgment by sentiment-word counting vs. SVM evaluated on human training sample

Training vs. test set size ratio	SVM training results							% in test set consistently judged by SVM and *Neg_net*
	Negative human-labeled news			Positive human-labeled news			Weighted	
	Precision	Recall	F1-score	Precision	Recall	F1-score	F1-score	
7:3	88.05%	88.40%	88.22%	88.35%	88.00%	88.18%	88.20%	83.60%
8:2	87.92%	88.80%	88.36%	88.69%	87.80%	88.24%	88.30%	85.40%
6:4	88.99%	86.50%	87.73%	86.87%	89.30%	88.07%	87.90%	83.80%

Table VII The Associations between Sentiment Measures and Returns

The dependent variables are industry- and size-adjusted returns over various horizons during days [-10, 10]. In Panel B, each row represents a regression similar to Panel A but uses a different sentiment measure. The control variable results in Panel B are omitted for brevity. All regressions include firm and individual month fixed effects. *T*-statistics are two-way cluster-adjusted at the firm and news-date levels, and are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Return association with *Neg_net*

	Industry- and size-adjusted return over day(s)								
	[-10, -6]	[-5, -3]	[-2]	[-1]	[0]	[1]	[2]	[3, 5]	[6, 10]
Neg_net	-1.212***	-1.605***	-2.171***	-6.787***	-11.366***	-1.432***	0.096	0.558***	0.153***
	(-18.40)	(-19.33)	(-16.87)	(-38.43)	(-50.78)	(-11.66)	(0.80)	(8.57)	(3.11)
beta	-0.035***	-0.010	-0.027	0.001	0.011	-0.011	-0.035*	-0.034***	-0.037***
	(-2.61)	(-0.65)	(-1.26)	(0.04)	(0.45)	(-0.58)	(-1.81)	(-2.74)	(-3.43)
Log market cap.	-0.231***	-0.211***	-0.207***	-0.237***	-0.225***	-0.128***	-0.088***	-0.065***	-0.083***
	(-15.03)	(-14.52)	(-9.98)	(-11.57)	(-10.57)	(-6.30)	(-4.81)	(-5.58)	(-7.77)
Book to market	0.260***	0.270***	0.266***	0.361***	0.321***	0.306***	0.233***	0.155***	0.130***
	(6.92)	(6.65)	(4.91)	(6.10)	(5.35)	(5.83)	(4.69)	(4.70)	(4.51)
Turnover	-4.964***	-3.328***	-3.310***	-3.400***	-3.682***	-2.483***	-2.175***	-1.849***	-1.512***
	(-16.07)	(-10.16)	(-7.19)	(-6.63)	(-7.62)	(-5.47)	(-4.94)	(-6.81)	(-7.04)
Volatility	4.815***	3.093***	3.225***	4.022***	1.784*	0.636	0.656	-0.047	-0.585
	(7.09)	(4.34)	(3.33)	(3.87)	(1.82)	(0.68)	(0.67)	(-0.07)	(-1.24)
SUE	0.008***	0.006*	0.012***	0.005	0.007	0.004	0.011***	0.006**	0.003
	(2.75)	(1.96)	(2.91)	(1.22)	(1.58)	(1.07)	(2.81)	(2.48)	(1.55)
Dividend yield	0.131	0.136	-0.162	-0.178	-0.362	-0.581	-0.528	-0.505	-1.023***
	(0.31)	(0.29)	(-0.23)	(-0.26)	(-0.46)	(-0.90)	(-0.84)	(-1.16)	(-2.88)
Stock age	-0.110***	-0.052**	-0.088**	-0.079*	-0.160***	-0.084**	-0.052	-0.004	-0.013
	(-4.61)	(-1.97)	(-2.16)	(-1.74)	(-3.66)	(-2.23)	(-1.39)	(-0.19)	(-0.70)
CSI300 dummy	-0.028*	-0.021	-0.042*	-0.015	-0.043	-0.058**	-0.041*	-0.037**	-0.029**
	(-1.69)	(-1.14)	(-1.70)	(-0.50)	(-1.50)	(-2.48)	(-1.80)	(-2.34)	(-2.17)
SOE dummy	0.017	-0.005	0.043	0.060	0.052	0.017	0.003	-0.038	0.004
	(0.51)	(-0.14)	(1.20)	(1.37)	(1.13)	(0.37)	(0.07)	(-1.41)	(0.17)
Excess Return$_{t-5,t-3}$			0.152***	0.062***	-0.045***	-0.037***	-0.028***	-0.006	-0.020***
			(16.61)	(6.94)	(-5.29)	(-4.85)	(-4.12)	(-1.41)	(-6.88)
Excess Return$_{t-10,t-6}$		0.059***	0.003	-0.016	-0.026***	-0.023***	-0.014	-0.026***	-0.027***
		(7.44)	(0.26)	(-1.51)	(-2.85)	(-2.58)	(-1.57)	(-5.61)	(-8.23)
Excess Return$_{m-12,m-2}$	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.000	-0.000	-0.000
	(-5.28)	(-4.86)	(-4.00)	(-3.20)	(-4.96)	(-3.39)	(-1.49)	(-0.88)	(-0.20)
Historical articles	0.060***	0.026***	-0.018	-0.045***	-0.067***	-0.037***	-0.026**	-0.018**	-0.016***
	(7.29)	(2.93)	(-1.57)	(-3.63)	(-5.11)	(-3.24)	(-2.47)	(-2.53)	(-2.82)
Number of articles$_t$	0.056***	0.117***	0.167***	0.320***	0.397***	0.003	-0.040***	-0.037***	-0.017***
	(6.50)	(10.36)	(9.61)	(15.41)	(15.85)	(0.16)	(-2.82)	(-4.43)	(-2.74)
Constant	5.583***	4.905***	5.136***	5.678***	5.793***	3.417***	2.380***	1.662***	2.132***
	(14.80)	(13.44)	(9.86)	(11.01)	(10.66)	(6.80)	(5.10)	(5.70)	(7.97)
Observations	413,156	411,751	410,943	411,751	411,751	410,145	408,985	408,703	407,763
Adj R-squared	0.031	0.027	0.025	0.032	0.050	0.013	0.011	0.016	0.022

Panel B: Return association with other sentiment measures

	Industry- and size-adjusted return over day(s)								
	[-10, -6]	[-5, -3]	[-2]	[-1]	[0]	[1]	[2]	[3, 5]	[6, 10]
Neg	-0.632***	-0.433***	-0.418	-4.942***	-12.916***	-1.382***	0.216	0.489***	-0.026
	(-5.12)	(-2.96)	(-1.61)	(-15.82)	(-37.94)	(-6.67)	(1.11)	(4.21)	(-0.28)
Pos	1.715***	2.459***	3.399***	8.874***	12.743***	1.724***	-0.046	-0.700***	-0.270***
	(20.39)	(22.31)	(21.81)	(41.00)	(45.65)	(10.79)	(-0.28)	(-8.69)	(-4.46)
PoliticalPos	0.035	0.058	0.775***	3.267***	6.837***	0.990***	-0.102	-0.306***	0.003
	(0.31)	(0.41)	(4.01)	(13.56)	(24.00)	(5.31)	(-0.36)	(-2.94)	(0.03)

Table VIII Return Regressions Adjusted for News Sentiment Persistence

Each row in this table reports the regression results of Table VII for a standardized abnormal sentiment measure. The control variable results are omitted for brevity. *T*-statistics are two-way cluster-adjusted at the firm and news-date levels, and are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Industry- and size-adjusted return over day(s)								
	[-10, -6]	[-5, -3]	[-2]	[-1]	[0]	[1]	[2]	[3, 5]	[6, 10]
Ab_Neg_net	0.006***	0.000	-0.006	-0.137***	-0.271***	0.001	0.033***	0.041***	0.020***
	(2.70)	(0.14)	(-1.23)	(-22.06)	(-36.48)	(0.29)	(6.75)	(17.03)	(10.47)
Ab_Neg	0.003	0.016***	0.030***	-0.037***	-0.200***	-0.008*	0.009**	0.019***	0.009***
	(1.39)	(6.03)	(5.89)	(-6.01)	(-30.30)	(-1.77)	(1.97)	(8.16)	(4.79)
Ab_Pos	-0.005**	0.008***	0.023***	0.139***	0.216***	-0.009*	-0.032***	-0.037***	-0.020***
	(-2.13)	(2.74)	(5.01)	(23.44)	(30.98)	(-1.94)	(-6.52)	(-14.81)	(-10.32)
Ab_PoliticalPos	-0.004*	-0.007**	0.003	0.054***	0.095***	0.011**	-0.009*	-0.012***	-0.007***
	(-1.88)	(-2.55)	(0.64)	(10.29)	(16.58)	(2.48)	(-1.74)	(-4.99)	(-4.01)

Table IX Horse Race Between Our and Alternative Dictionaries

This table reports the return regression of Table VII, with sentiment measures from the four dictionaries pooled as independent variables. The control variable results are omitted for brevity. *T*-statistics are two-way cluster-adjusted at the firm and news-date levels, and are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Industry- and size-adjusted return over day(s)								
	[-10, -6]	[-5, -3]	[-2]	[-1]	[0]	[1]	[2]	[3, 5]	[6, 10]
Neg_net	-1.010***	-1.394***	-1.725***	-5.591***	-10.645***	-1.197***	-0.144	0.543***	0.108
	(-10.35)	(-11.40)	(-8.95)	(-24.33)	(-37.87)	(-6.73)	(-0.78)	(5.18)	(1.45)
Neg_net_YZZ	-0.345***	-0.591***	-1.658***	-3.508***	-2.683***	-0.565***	0.122	0.171*	0.161**
	(-3.80)	(-5.35)	(-9.00)	(-15.13)	(-12.90)	(-3.36)	(0.71)	(1.73)	(2.21)
Neg_net_LM	0.158	0.694***	2.661***	4.921***	3.910***	0.817***	0.334	-0.343***	-0.220**
	(1.27)	(4.59)	(10.33)	(14.45)	(13.61)	(3.72)	(1.40)	(-2.97)	(-2.32)
Neg_net_generic	1.327***	1.959***	2.752***	5.498***	8.302***	-0.524	0.524	-0.479	-0.570**
	(3.99)	(4.90)	(4.59)	(8.39)	(10.94)	(-0.89)	(0.71)	(-1.44)	(-2.27)

Table X Comparison with the YZZ Dictionary

In Panel A, we run the return regression of Table VII with *Neg_net_2500* (the net negative tone based on the dictionary constructed from reading 2,500 articles) and *Neg_net_YZZ* both in the regression. In Panel B, we break the sample into agreeing and disagreeing news, where agreeing news refers to the news articles with the tone measure appearing simultaneously as below (or above) the median tone value using both the sentiment measures from the dictionary of reading 2,500 articles and the YZZ dictionary. The control variable results are omitted for brevity. *T*-statistics are two-way cluster-adjusted at the firm and news-date levels, and are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Panel A: *Neg_net_2500* controlled for *Neg_net_YZZ*

	Industry- and size-adjusted return over day(s)								
	[-10, -6]	[-5, -3]	[-2]	[-1]	[0]	[1]	[2]	[3, 5]	[6, 10]
*Neg_net_*2500	-1.006***	-1.331***	-1.472***	-5.132***	-10.341***	-1.113***	-0.100	0.509***	0.090
	(-10.25)	(-11.02)	(-7.50)	(-21.42)	(-37.38)	(-6.31)	(-0.55)	(4.89)	(1.22)
Neg_net_YZZ	-0.237***	-0.312***	-0.800***	-1.898***	-1.197***	-0.366**	0.228	0.055	0.072
	(-2.77)	(-2.92)	(-4.93)	(-10.06)	(-6.27)	(-2.35)	(1.28)	(0.60)	(1.07)

Panel B: Agreeing vs. disagreeing of news with YZZ

	Industry- and size-adjusted return over day(s)									
	Agreeing News					Disagreeing News				
	[-2]	[-1]	[0]	[1]	[2]	[-2]	[-1]	[0]	[1]	[2]
Net negative tone										
*Neg_net_*2500	-1.290***	-5.695***	-11.350***	-1.417***	0.102	-1.838***	-1.377**	-9.150***	0.119	0.064
	(-5.58)	(-20.01)	(-35.59)	(-6.87)	(0.47)	(-3.32)	(-2.28)	(-13.09)	(0.22)	(0.12)
Neg_net_YZZ	-0.955***	-1.363***	-0.260	-0.106	0.037	-0.626	-0.960*	-3.833***	-0.333	1.134**
	(-4.92)	(-6.04)	(-1.13)	(-0.57)	(0.17)	(-1.40)	(-1.80)	(-7.39)	(-0.72)	(2.53)
Obs.	352,913	353,628	353,628	352,197	351,193	57,915	58,009	58,009	57,834	57,678
Negative tone										
*Neg_*2500	1.928***	-1.474***	-8.525***	-1.062***	0.082	0.089	-1.852**	-4.647***	0.707	-0.376
	(4.66)	(-3.12)	(-18.31)	(-3.31)	(0.25)	(0.12)	(-2.20)	(-5.41)	(0.92)	(-0.50)
Neg_YZZ	-2.887***	-4.613***	-6.364***	-0.611**	0.290	-2.037***	-3.697***	-7.117***	-1.085*	0.133
	(-8.62)	(-13.49)	(-16.83)	(-2.09)	(0.99)	(-3.57)	(-5.84)	(-11.68)	(-1.93)	(0.24)
Positive tone										
*Pos_*2500	3.821***	9.383***	15.116***	1.794***	-0.146	1.858***	2.337***	9.745***	0.465	0.981
	(12.63)	(25.46)	(33.44)	(6.24)	(-0.50)	(3.19)	(3.74)	(12.96)	(0.77)	(1.60)
Pos_YZZ	-0.403	-0.117	-2.059***	0.005	0.002	-0.100	2.121***	3.904***	0.452	0.183
	(-1.53)	(-0.36)	(-5.87)	(0.02)	(0.01)	(-0.18)	(3.47)	(5.70)	(0.77)	(0.33)

Table XI Sentiment Bias in State Media

The dependent variables are either the sentiment measures, or *PoliticalNouns* (the fraction of total count of political phrases in the Dictionary of Scientific Development relative to the total number of words in a news article), each multiplied by 100. *MediabiasIndex* equals *PoliticalPos* minus *Neg*. State media is a dummy variable that equals one if the article is from a media outlet whose ultimate control right belongs to the central or a provincial government. All regressions include firm and individual month fixed effects. *T*-statistics are two-way cluster-adjusted at the firm and news-date levels, and are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	(1)	(2)	(3)	(4)
	PoliticalPos	*Neg*	*MediabiasIndex*	*PoliticalNouns*
State media	0.258***	-0.271***	0.529***	0.615***
	(9.68)	(-11.85)	(12.28)	(18.59)
beta	-0.129***	0.217***	-0.345***	0.112***
	(-5.08)	(6.44)	(-7.03)	(3.06)
Log market cap.	0.076**	0.003	0.073	-0.179***
	(2.32)	(0.08)	(1.26)	(-5.02)
Book to market	-0.256***	0.728***	-0.984***	-0.594***
	(-3.13)	(6.86)	(-6.06)	(-6.10)
Turnover	-0.062	-0.275	0.213	0.965
	(-0.13)	(-0.51)	(0.25)	(1.60)
Volatility	-1.500	4.835***	-6.335***	-4.605***
	(-1.57)	(4.24)	(-3.78)	(-3.89)
SUE	0.046***	-0.072***	0.118***	0.048***
	(6.50)	(-8.32)	(8.50)	(4.99)
Dividend yield	-1.459	-1.791	0.332	1.538
	(-1.22)	(-1.47)	(0.16)	(1.53)
Stock age	-0.137***	0.163***	-0.300***	0.146***
	(-3.25)	(2.70)	(-3.91)	(2.67)
CSI300 dummy	-0.022	0.022	-0.044	-0.042
	(-0.42)	(0.29)	(-0.41)	(-0.65)
SOE dummy	-0.118	0.055	-0.173	0.037
	(-0.87)	(0.34)	(-0.61)	(0.43)
Histotical articles	-0.073***	0.198***	-0.271***	-0.111***
	(-3.45)	(7.10)	(-6.57)	(-4.12)
Excess Return$_{t-1}$	2.043***	-3.337***	5.380***	0.610**
	(11.67)	(-13.92)	(16.25)	(2.10)
Excess Return$_{t-2}$	0.019	-0.108	0.127	1.803***
	(0.12)	(-0.51)	(0.43)	(7.76)
Excess Return$_{t-5,t-}$	-0.229	-0.568	0.339	-0.193
	(-0.69)	(-1.47)	(0.58)	(-0.49)
Excess Return$_{t-}$	-0.214	-1.999***	1.785**	-1.227**
	(-0.49)	(-4.22)	(2.42)	(-2.23)
Excess Return$_{m-}$	0.076***	-0.100***	0.177***	0.007
	(4.89)	(-5.48)	(6.13)	(0.36)
Constant	2.382***	0.456	1.925	9.091***
	(3.06)	(0.52)	(1.35)	(10.24)
Observations	657,649	657,649	657,649	657,649
Adj R-squared	0.103	0.093	0.124	0.143

Table XII State Media Sentiment Bias and Returns

State media is a dummy variable that equals one if the article is from a media outlet whose ultimate control right belongs to the central or a provincial government. *MediabiasIndex* equals *PoliticalPos* minus *Neg*. All regressions include firm and individual month fixed effects. *T*-statistics are two-way cluster-adjusted at the firm and news-date levels, and are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Industry- and size-adjusted return over day(s)							
	[-1, 1]	[-1, 1]	[-1, 1]	[-1, 1]	[-1, 1]	[-1, 1]	[-1, 1]	[-2, 2]
State media	-0.142*** (-8.43)	-0.381*** (-17.24)	-0.221*** (-15.76)	-0.172*** (-9.48)	-0.142*** (-7.20)	-0.426*** (-15.86)	-0.274*** (-13.84)	-0.191*** (-12.96)
PoliticalPos	5.154*** (11.79)				5.189*** (11.68)			
State media × *PoliticalPos*	-2.187*** (-5.77)				-2.217*** (-5.77)			
Neg		-12.041*** (-20.10)				-12.150*** (-20.14)		
State media × *Neg*		5.497*** (11.74)				5.595*** (11.87)		
MediabiasIndex			7.070*** (17.86)				7.217*** (17.91)	4.820*** (16.33)
State media × *MediabiasIndex*			-3.108*** (-10.30)				-3.248*** (-10.55)	-2.350*** (-10.32)
PoliticalNouns				0.495* (1.84)	-0.204 (-0.74)	-0.844*** (-2.95)	-1.270*** (-4.36)	-0.711*** (-3.18)
State media × *PoliticalNouns*				-0.579** (-2.17)	0.055 (0.20)	0.948*** (3.31)	1.211*** (4.17)	0.994*** (4.38)
beta	-0.012 (-0.45)	0.002 (0.08)	0.001 (0.04)	-0.016 (-0.64)	-0.011 (-0.45)	0.002 (0.08)	0.001 (0.06)	-0.011 (-0.51)
Log market cap.	-0.265*** (-7.96)	-0.263*** (-8.10)	-0.266*** (-8.10)	-0.263*** (-7.90)	-0.266*** (-7.97)	-0.263*** (-8.09)	-0.266*** (-8.10)	-0.234*** (-7.93)
Book to market	0.290*** (3.55)	0.340*** (4.28)	0.330*** (4.09)	0.280*** (3.42)	0.289*** (3.53)	0.340*** (4.27)	0.329*** (4.08)	0.301*** (4.35)
Turnover	-3.569*** (-7.14)	-3.602*** (-7.35)	-3.583*** (-7.29)	-3.575*** (-7.11)	-3.568*** (-7.13)	-3.602*** (-7.34)	-3.581*** (-7.28)	-3.253*** (-7.67)
Volatility	3.489*** (2.92)	3.852*** (3.28)	3.771*** (3.19)	3.421*** (2.86)	3.483*** (2.92)	3.852*** (3.28)	3.764*** (3.19)	2.905*** (3.02)
SUE	0.007 (1.33)	0.003 (0.61)	0.003 (0.61)	0.009 (1.62)	0.008 (1.34)	0.003 (0.59)	0.003 (0.61)	0.006 (1.36)
Dividend yield	0.003 (0.00)	-0.182 (-0.18)	-0.039 (-0.04)	-0.076 (-0.07)	0.006 (0.01)	-0.170 (-0.17)	-0.020 (-0.02)	-0.190 (-0.22)
Stock age	-0.153*** (-3.67)	-0.147*** (-3.69)	-0.146*** (-3.59)	-0.157*** (-3.77)	-0.153*** (-3.66)	-0.146*** (-3.67)	-0.144*** (-3.56)	-0.129*** (-3.80)
CSI300 dummy	-0.129*** (-3.38)	-0.129*** (-3.32)	-0.128*** (-3.29)	-0.131*** (-3.43)	-0.130*** (-3.38)	-0.130*** (-3.33)	-0.129*** (-3.29)	-0.096*** (-2.83)
SOE dummy	0.005 (0.10)	0.005 (0.11)	0.010 (0.20)	-0.001 (-0.02)	0.005 (0.10)	0.006 (0.12)	0.011 (0.23)	0.004 (0.09)
Excess Return$_{t-5,t-3}$	2.365*** (3.15)	2.280*** (3.06)	2.316*** (3.10)	2.364*** (3.14)	2.365*** (3.15)	2.278*** (3.06)	2.313*** (3.10)	5.022*** (7.77)
Excess Return$_{t-10,t-6}$	0.344 (0.37)	0.178 (0.20)	0.251 (0.27)	0.336 (0.37)	0.342 (0.37)	0.178 (0.20)	0.249 (0.27)	0.235 (0.31)
Excess Return$_{m-12,m-2}$	-0.050*** (-2.92)	-0.055*** (-3.26)	-0.056*** (-3.29)	-0.047*** (-2.75)	-0.050*** (-2.92)	-0.055*** (-3.26)	-0.056*** (-3.29)	-0.051*** (-3.58)
Historical articles	-0.077*** (-3.48)	-0.062*** (-2.84)	-0.066*** (-2.97)	-0.080*** (-3.58)	-0.077*** (-3.49)	-0.063*** (-2.86)	-0.067*** (-3.02)	-0.057*** (-3.22)
Constant	7.474*** (9.06)	7.785*** (9.65)	7.514*** (9.25)	7.600*** (9.20)	7.489*** (9.09)	7.821*** (9.69)	7.574*** (9.32)	6.586*** (9.01)
Observations	659,299	659,299	659,299	659,299	659,299	659,299	659,299	659,299
Adj. R-squared	0.068	0.078	0.076	0.065	0.068	0.078	0.076	0.088

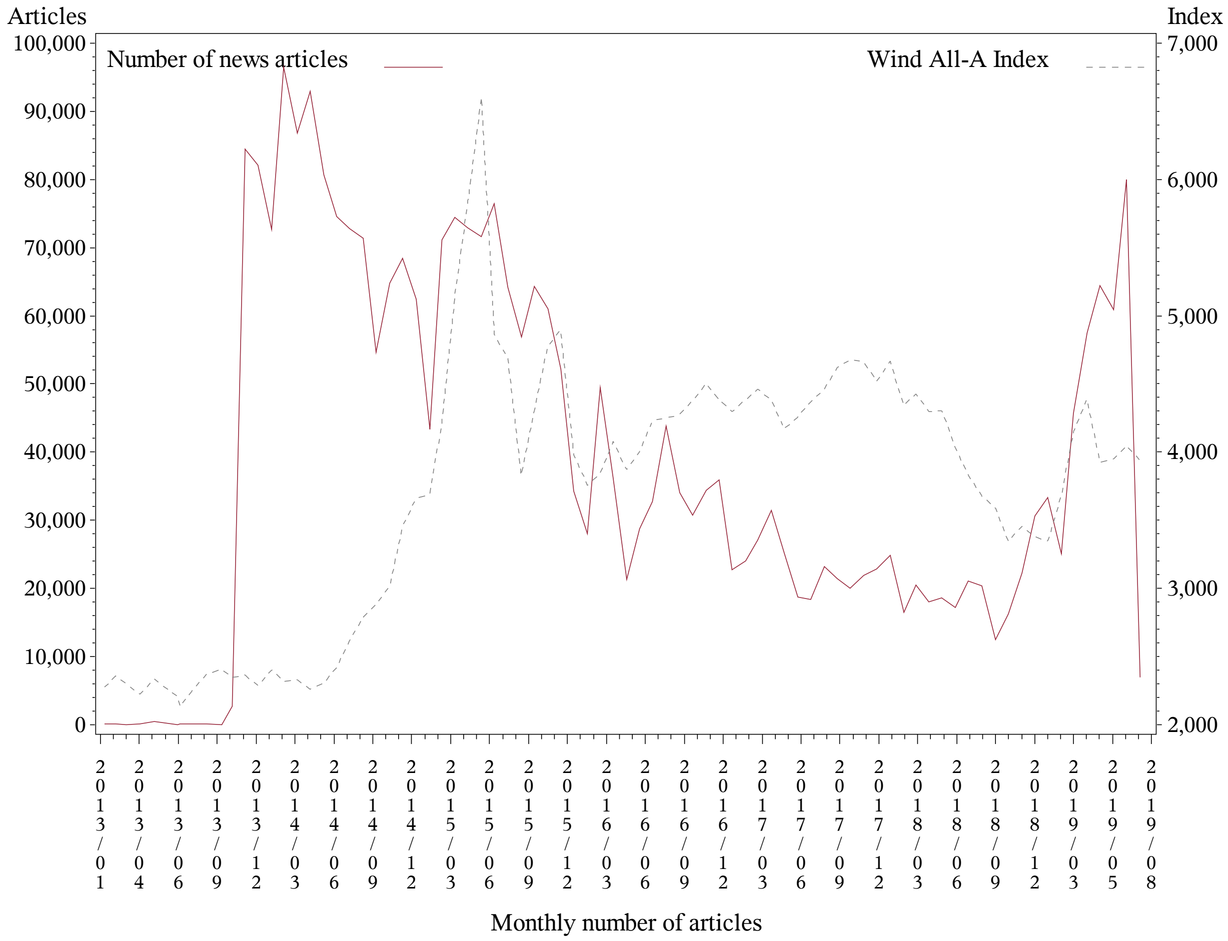


Figure 1: The solid line shows the monthly number of news articles used in Word2vec. For August 2019 (the last month of our sample period), we download articles for only part of the month. The gray dotted line shows the monthly average closing level of the Wind All-A stock index.

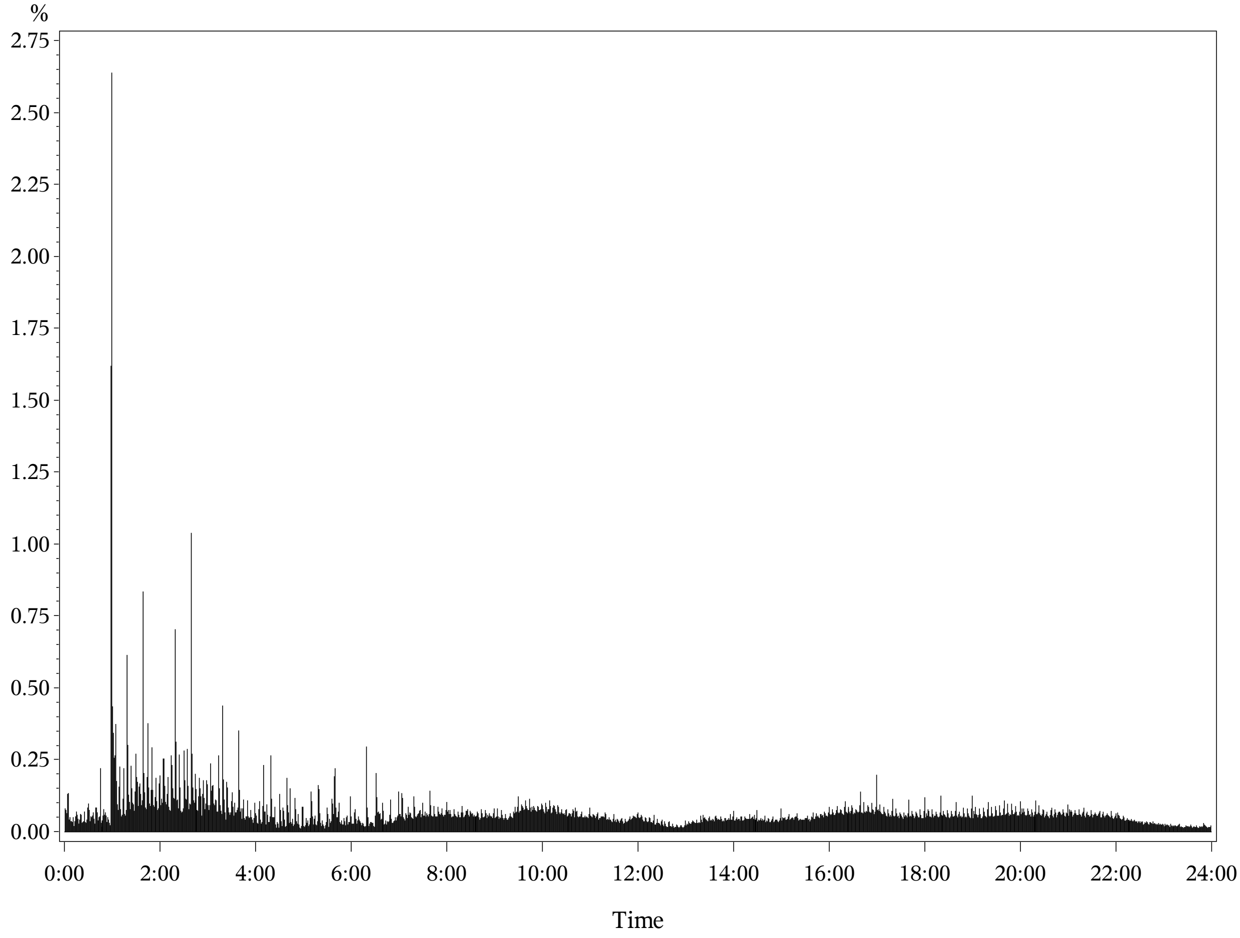


Figure 2: This figure shows the percentage histogram of the time for the news articles released during a trading day.

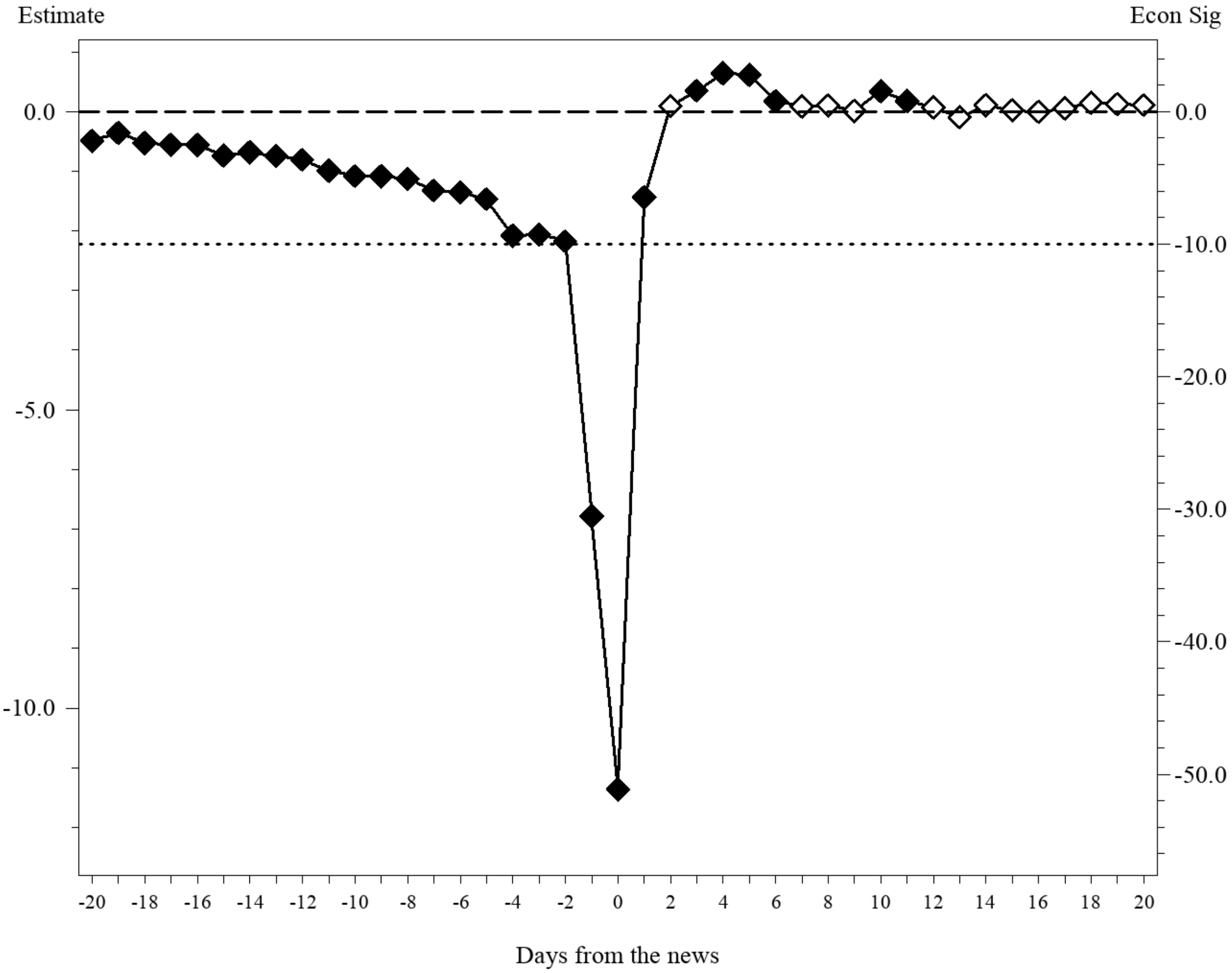


Figure 3: The magnitudes of the coefficient estimates and economic significance of *Neg_net* in daily return regressions from days [-20] to [20]. We run a regression for each day's excess return. For days [-2] to [2], the regression specification follows Table VII. For other days, the regression covariates include those in Table VII except the control variables of the short-term excess returns for days [-10, -6] and [-5, -3]. Solid (blank) diamond indicates (in)significance at the 10% level. For the right axis, "Econ Sig" refers to the coefficient estimate times the standard deviation of *Neg_net*, multiplied by 100 to be expressed in bps.

Internet Appendix to
"Language and Domain Specificity:
A Chinese Financial Sentiment Dictionary"

This Internet Appendix reports results supplementary to the paper "*Language and Domain Specificity: A Chinese Financial Sentiment Dictionary*." We compare the tokenization results of three common word segmenters, provide the list of news sources for our sample, describe the implementation of Word2vec and SVM, conduct the word cloud analysis for our dictionary, and present additional return regression results.

I. Tokenization Comparison of Three Segmenters

There are a number of word segmenters that are frequently used for Chinese to break sentences into word phrases, or computational linguistic "tokens." We choose the following three tokenization packages: Jieba, which is perhaps the most frequently used Chinese segmenter, Stanford Word Segmenter developed by Stanford University, and NLPIR-ICTCLAS developed by Chinese Academy of Sciences. We randomly choose four sentences from financial news on, respectively, stock market commentary, share repurchase, real estate, and firm litigation. We benchmark the tokenization results of the segmentation packages with that read by human. Table A1 shows that among the four example sentences, Jieba is by far the best performing package, offering an almost perfect segmentation in two sentences, and the least inconsistency with human reading for the remaining two sentences. In these examples, the outperformance in Jieba clearly lies in its ability to correctly identify phrases related to finance and economics; for example, both Stanford Segmenter and NLPIR-ICTCLAS are not able to correctly identify the phrase "black swan" (a common finance term in Chinese used to refer to an unpredicted and impactful negative event) and "price increase limit" (a term indicating that the stock price increase hits the designated daily limit), both of which appear in our sentiment-word list and are correctly tokenized by Jieba. Stanford Segmenter and NLPIR-ICTCLAS treat "black swan" as two words of "black" plus "swan", none of which appears on our sentiment-word list. Untabulated, we also expand the exercise to 100 sentences and find that the results are similar. These results suggest that Jieba is suitable for our purposes of sentiment dictionary construction.

[Insert Table A1 about here.]

II. List of News Sources

Table A2 lists the 64 news sources that each supply at least 1,000 articles for our full sample used for dictionary construction. All these sources belong to the main-stream media. Chinese Securities Daily, which is owned and operated by Xinhua News Agency, supplies the largest number of articles, accounting for one seventh of the full sample, and one quarter of the firm-specific news sample used in return and media bias analyses. The four designated media outlets by China Securities Regulatory Commission (the Chinese SEC) where firms publish public announcements and regulatory filings (China Securities Journal, Shanghai Securities News, Securities Times, and Securities Daily) are all in the top six supplier list, accounting for about half of the articles.

[Insert Table A2 about here.]

III. Parameterization and Implementation of Word2vec and SVM

In this section we provide technical details in implementing Word2vec and SVM in this paper. Following Mikolov et al. (2013), among many others, we apply the Skip-gram model, which predicts the closest context words (surrounding words) given a target word. The skip-gram model can be regarded as a simple neural network model with three layers—input, projection, and output. We first encode a target word into a one-hot vector, with the value of one for the position of target word and zero for all other words in the vocabulary consisting of the sample corpus. The one-hot vector is the input layer; the dimension of the input layer is thus the number of the unique words (V) in our corpus. In the projection layer there is one N-dimensional hidden layer, where N is the number of hidden "features" of word vectors that characterize the word vector. The input layer is projected to the hidden layer through a weight matrix (W) with a dimension $V \times N$ featuring the N dimensions for each of the word in V. The output layer is a softmax regression classifier that obtains the posterior distribution of words given the target word. The training objective in the output layer of the skip-gram model is to minimize the summed prediction error across all context words of the target word within a given window by learning the optimal weight matrix W, or the word embeddings. This optimal matrix allows us to calculate the cosine similarity between the target word and other words, and we then pick the ones with the highest cosine similarities to be the target word's semantically close words. We use Python's "genism" package to implement the Word2vec model. Our resulting number of unique words (i.e., the value of V) is 72,092, which is about two thirds of 106,230 unique words commonly perceived in modern Chinese.[1] We set *N* to 250, require training words to appear in our corpus at least 100 times, and examine the context words respectively, for window sizes of 2, 3, 5, and 7 for only complete sentences that contain seven or more words.

As the size of our corpus becomes large in later iterations (see Table 1), we use two methods suggested in Mikolov et al. (2013) to increase computing efficiency and improve accuracy. The first is to randomly "subsample" frequent words, i.e., randomly remove words with low frequency from the vector *V*. The second is negative sampling that forms a "negative" sample using part of the words outside the context window (which is our "positive" sample) rather than all of the words. See Mikolov et al. (2013) for subsampling and negative sampling details. We set the subsampling threshold to 10^{-5} (the cutoff for word removal, defined as the number of a word's appearances to total number of words in the corpus) and select 5 negative samples for each of our target word training. In addition, we execute our full sample training of 3.08 million news articles with the parallel computing tool Apache Spark3.

Word2vec is an unsupervised machine learning process. We manually filter the output of Word2vec. In particular, we select our sentiment word from seven words that have the highest cosine similarity to the target word. For example, for the positive seed word "涨停" (verb of "price hitting up-limit"), Word2vec produces the following top seven candidates: "涨停板" (noun of "price hitting up-limit"), "跌停" (verb of "price hitting down-limit", which is its antonym), "一字板" (another noun of "price hitting up-limit"), "封板" (another verb of "price hitting up-limit"), "大涨" ("stock price soars"), "拉升" ("stock price gap-increases"), and "两连板" ("two continuous hits of price up-limit"). We proofread these candidate terms into positive, negative, political, or neutral, which in this case, six terms are labeled as positive-sentiment words and the antonym as a negative-sentiment word.

We perform the article-level sentiment classification using the more traditional supervised machine method of support vector machine (SVM). SVM is a classifier that divides the sample into distinct groups by finding a hyperplane to maximize the distance (i.e., distinctiveness) among the groups (e.g., Cortes and Vapnik, 1995; Vapnik, 2013). We represent each news article by the

[1] See, e.g., https://www.tutormandarin.net/en/how-many-chinese-words-are-there-how-many-chinese-characters/.

TF-IDF (term frequency-inverse document frequency) vector, which is a matrix consisting of the frequency of a term appearing in a given document, adjusted by the number of documents that contain the word to accommodate for the fact that some words appear more frequently in writing in general. The distance between two vectors is calculated as their dot products (which is called "kernel function"), and we use both linear kernel and nonlinear kernel of radial basis function (RBF). We implement 10-fold cross-validation (e.g., Kohavi, 1995), where the training set is randomly partitioned into 10 subsamples with equal size. Then, one of the 10 subsamples is retained as the validation set, and the remaining nine subsamples are used for the SVM training. After repeating the validation 10 times, with each subsample treated as the validation set, the average performance of each fold (i.e., each iteration) aggregates to the overall performance of our SVM training. During the process, two governing hyperparameters that are used to address misclassification of the SVM—the penalty for each misclassified data point and the distance of influence of a single training point (the latter only in the case of RBF)—are decided by random search and then applied to the final model training of the entire training set (Bergstra and Bengio, 2012). We use the Python package "scikit-learn" to conduct the SVM training.

IV. Word Cloud Analysis of our Dictionary

Figure A1 shows the word clouds of our top 50 negative, positive, and political words in the sample of 3.1 million news articles. Word clouds visualize the frequency of appearance of each word by its relative size. The top three negative words are "波动" ("volatile"), "减持" ("shareholding reduction"), and "异常" ("abnormality"); the top three positive words are "超过" ("exceed"), "发展" ("develop"), and "同意" ("approve"); and the top three political words are "保证" ("guarantee"), "维护" ("maintain"), and "推进" ("advance"). The negative-word list includes some uncertainty words, as uncertainty words that appear in financial news often carry pessimistic connotation; in a concurrent dictionary, Yao et al.'s (2021) list of negative words also include uncertainty words.

Within each category of the top 50 words, we mark those appearing in the three alternative dictionaries (LM, YZZ, and Generic) to be gray, and the original words in our dictionary to be bold. In the top-50 words, we add 8 negative, 13 positive, and 21 political words to the existing dictionaries. The fact that we contribute the most to the top political words among the three sentiment categories is consistent with our political word list being original in the literature.

[Insert Figure A1 about here.]

V. Robustness Tests of Sentiment's Return Associations

5.1 Portfolio sorting

To corroborate the economic significance of returns to sentiment, Table A3 examines the univariate association between news article sentiment and the day [0] and [1] returns by sorting stocks on sentiment measures. In Panel A, we sort the sample into decile portfolios on the ranked value of *Neg_net*, and then calculate the equal-weighted returns on these decile portfolios. Day [0] (Day [1]) portfolio returns are (almost) monotonically decreasing in the rank value of *Neg_net*. The return difference between deciles 1 and 10 is 2.11% for Day [0] and 0.25% for Day [1], both highly statistically significant. Moreover, the majority of these hedge portfolio returns come from the long leg of the portfolio (in that the positive side of returns in decile 1 dominates the negative side of returns in decile 10). As there is only a limited degree of short-selling in the Chinese market, it is unlikely that limits to arbitrage such as inability to short would drive the existence of these hedge portfolio returns.[2] These results suggest that *Neg_net* negatively predicts day [0] and [1] returns; the effect is particularly strong on day [0].

[Insert Table A3 about here.]

The remainder of Panel A shows the alphas of these decile portfolios. We adjust raw returns by beta calculated from the past three years' monthly returns of the stock using the traditional Capital Asset Pricing Model with the market return proxied by the Wind All-A Index. By and large, the alphas of these portfolios exhibit similar patterns as those of the raw returns. The alphas of the decile 1 over 10 hedge portfolios are 1.99% and 0.34%, respectively, for days [0] and [1], similar in magnitude to the unadjusted returns. The return differences of the hedge portfolios, therefore, do not seem to be driven by differential beta risks of the underlying stocks.

[2] Short-selling is only allowed in China after 2010 and is heavily regulated. As a point of illustration, Reuters reports that on June 16, 2020 the ratio of short interest to total market capitalization in China is 0.044%, as opposed to 2.4% in the US (https://www.reuters.com/article/uk-china-market-shortselling-analysis-idUSKBN23P3TJ).

The univariate return patterns of *Neg_net* extend to other sentiment measures. In Panel B, we present the days [0] and [1] alphas of portfolios sorted on *Neg*, *Pos*, and *PoliticalPos*, respectively. As expected, *Neg* is negatively related to alphas, while *Pos* and *PoliticalPos* are positively related to alphas. The positive return relation of *PoliticalPos* is consistent with the fact that the Chinese stock market is heavily influenced by government policies. The decile 1 over decile 10 hedge portfolio alphas on days [0] and [1] across *Neg* and *Pos* are both about two thirds of the magnitude of their *Neg_net* counterparts. Contrary to the literature that emphasizes the predictive return power of *Neg* in the English-language setting (e.g., Tetlock, Saar-Tsechansky, and Macskassy, 2008), the Chinese *Pos* not only predicts days [0] and [1] alphas, but also carries hedged-alpha magnitudes greater than those of *Neg*. Lastly, while *PoliticalPos* also predicts alphas, its hedged-alpha magnitudes are significantly smaller than those of *Pos*. In sum, Table A3 corroborates that our news sentiment measures are associated with day [0] returns with a strong economic magnitude, and that both the negative and positive sides of the tone predict returns.

5.2 Short-term interaction between news and returns

In Table VII of the main text we report that *Neg_net* is significantly related to returns during days [-2, -1] with an economic significance above 10 bps. In Section 3.2 of the main text we show that using abnormal *Neg_net* by adjusting for the firm's own historical *Neg_net* value does not fully explain away the phenomenon of returns leading news sentiment. In this section we examine the possibility of short-term interaction between news and returns as an explanation. That is, news and returns could "iterate" on each other over a short period of time, in that, for example, news coverage drives returns, which, in turn, drives further news.

To address the short-term interaction between news and returns, we follow Huang, Tan, and Wermers (2020, "HTW") to group closely spaced news articles spanning multiple days into a "cluster," and treats all of the news within the cluster as a single event by averaging the news sentiment and return within the cluster and using the days around the cluster to bound the event. By design, the interactions between news and returns are suppressed within a news cluster. We follow HTW in isolating news strings by using three days as the stopping interval to cluster our news, that is, a particular firm will continue to have a single news-cluster event until news coverage stops for at least three trading days.

Panel A of Table A4 repeats the return regression results (as in Table VII). There are 199,370 news clusters, or roughly half of the news days in Table VII. Although the number of observations drops by half, we continue to observe similar patterns as in Table VII—negative (positive) coefficient significance of negative (positive) sentiment measures on returns measured over various windows during days [-10, 0], and the reversal of such a relation over windows during days [2, 10]. Importantly and consistently, days [0] and [-1] have a much larger economic significance than exhibited by other days, similar to Table VII. A similar calculation shows that day [0] has about an economic significance of −50 bps, and day [-1] has a significance of −22 bps, which are in the same order of magnitude with, respectively, −51 bps and −31 bps in Table VII. Untabulated, we also execute a news-clustering exercise for abnormal sentiment measures. We find, similarly, that only days [0] and [-1] exhibit considerably larger economic return significance than other days. The robustness test in Panel A thus indicates that various adjustments to reduce the effect of news persistence does not rule out the economic significance of the effect of sentiment on returns on day [-1]. This finding provides further evidence that there exists some degree of news leakage in China.

[Insert Table A4 about here.]

5.3 Firm-initiated vs. press-initiated news

Extant literature (e.g., HTW) suggests that there exist systematic differences between news released by a firm itself (“firm-initiated”) and news written by the press (“press-initiated”). Press-initiated news typically exercises editorial control over the news content, and can add important interpretations, relative to firm-initiated news (Bushee et al., 2010). Since firms tend to withhold negative information (e.g., Kothari, Shu, and Wysocki, 2009), firm-initiated news may be optimistically biased, or, at best, noisy versions of the “truth,” and thus, may exhibit lower return implications.

In the Chinese market, there are four designated media outlets through which firms publish public announcements and regulatory filings, as previously discussed. It is rare for firms to use sources outside of these designated outlets for announcements, for either regulatory-compliance or cost-saving purposes. In addition, different from firm-initiated news, a press-initiated news article almost always shows the name of the author with the keyword “Author”, “Journalist”, or “Editor” appearing at the beginning or end of the article. Utilizing these characteristics, we classify news that appears on the designated list without any of the author keywords as firm-initiated, and the rest as press-initiated. Firm-initiated news accounts for 48.1% of the news articles in our regression sample.

Using either press- or firm-initiated news, we re-calculate our news sentiment metrics for each. Panels B and C of Table A4 present the return regressions results for our *Neg_net* measure of sentiment, partitioned by press- vs. firm-initiated news. We find that, for both types of news, the results are similar to prior conclusions using the full sample of news; that is, meaningful return associations with sentiment take place on days [-1] and [0] for both types of news. Consistent with the idea that returns are less sensitive to firm-initiated news, we find that the sensitivity of returns to *Neg_net* of press-initiated news is larger than that of firm-initiated news in China. In untabulated results, we find that the difference is statistically significant. Thus, the return associations that we find are more pronounced in press-initiated news.

5.4 Considering only news headlines

News headlines summarize news articles. Professional terminals such as Wind and Bloomberg offer news feeds in headline streaming. It is, therefore, possible that stock prices incorporate headline information sooner than the body corpus of news. We calculate the sentiment measures for headlines, then repeat our Table VII regressions of returns. Panel D of Table A4 presents the results. We, again, find that the signs and economic magnitudes of the headline sentiment measures are consistent with those in Table VII, and that meaningful return associations concentrate on days [-1, 0]. For instance, for day [0], the headline *Neg_net* has a coefficient estimate of −4.266, amounting to an economic significance of −47.4 bps (= −4.266 × standard deviation of 0.1110).

5.5 Removing all intra-day news

Earlier, we assign all intra-trading-day news to the trading day as day [0] news. One potential limitation with this assignment may be that intra-day news experiences less than a full day of resultant trading, leading to a partial-day issue. Figure 2 earlier showed that about one sixth of our overall sample of news articles are released intra-day. We address the partial-day problem by removing these intra-day news articles. Panel E of Table A4 shows that the coefficient estimate of day [0] experiences a significant drop while the coefficient estimate of day [-1] remains similar—consistent with day [0] most keenly experiencing the partial-day issue; nevertheless, the day [0] coefficient remains highly statistically significant. The limited degree of information leakage that we identified in Section 3.2 also helps to explain this phenomenon. In the existence of information leakage, the impact of intra-day news might be most strongly reflected on the day before the news release (which includes the same-trading day time until the news is released)—which is not found in these results. Removing intra-day news, therefore, only reduces the magnitude of the sentiment coefficient estimate for day [0] returns. Overall, our main conclusions remain unchanged when we remove intra-day news.

5.6 Excluding days around earnings announcements

The literature suggests that stock prices react to earnings surprises prior to earnings announcements by exploiting, for example, sell-side analysts' slow movement towards the actual earnings number (e.g., Baker et al., 2010). News around earnings announcements consists of a large portion of news articles in the U.S. (e.g., HTW). In our sample, news articles released [-3, 3] days around an earnings announcement accounts for 14.4% of all Chinese news articles. Panel F of Table A4 removes the articles around earnings announcements and repeats the return regressions. Our results are robust in both statistical and economic significance.

5.7 Standalone return associations of the alternative dictionaries

Table A5 replicates the return regressions of Table VII of the main text for three alternative dictionaries: the LM dictionary ("LM dictionary"), the You, Zhang, and Zhang (2018) dictionary ("YZZ dictionary"), and the common set of three general dictionaries ("Generic dictionary"). Similar to our dictionary, we observe that the negative (positive) sentiment measures are generally statistically negatively (positively) related to returns for windows during days [-10, 1], and that the return effect is largely reversed for windows during days [2] to [10]. As expected, the Generic dictionary, which does not focus on the sentiment of words in the context of financial markets, has smaller statistical significance. As with our dictionary, the strongest association for all three dictionaries still occurs during days [-1, 0], where the magnitudes of coefficient estimates are much larger than those of other days.

[Insert Table A5 about here.]

Further, we find that the economic significance of the sentiment measures in the LM and Generic dictionaries is much smaller than that of our dictionary. *Neg_net_LM* (*Neg_net_generic*) has an economic significance of −23.2 (−5.0) bps for day [0] returns, and −10.5 (−3.6) bps for day [-1] returns, as compared to −51 and −31 bps for our dictionary's *Neg_net* (reported in the main text), respectively. That is, *Neg_net_LM* has an economic significance that is less than half, compared to that of our dictionary, for day [-1] and day [0] returns. The negative and positive sides of sentiment for the LM dictionary (*Neg_LM* and *Pos_LM*) compare relatively similarly with our dictionary's (*Neg* and *Pos*, respectively) results, and are omitted for brevity. *Neg_net_YZZ* has a smaller-yet-comparable economic significance (relative to *Neg_net*) of −44.1 bps on day [0] and −28.3 bps on day [-1]; and *Neg_YZZ* and *Pos_YZZ*'s economic significance compares relatively similarly to that of *Neg* and *Pos*.

5.8 Non-uniqueness of PoliticalPos in return regressions

One of the distinctive features of our dictionary is that we separate out politically-inclined positive words. We manually check the YZZ dictionary list of positive words, and find that it contains a significant number of political words, as defined by our dictionary. As shown in Table II of the main text, 14.5% of our political words are covered in the YZZ dictionary. In Table VII of the main text, we also showed that the return associations of *PoliticalPos* are weaker than other types of words. A natural ensuing question is then whether *PoliticalPos* is incrementally related to returns, given the presence of *Pos* or *Pos_YZZ*.

Panel A of Table A6 includes both *PoliticalPos* and *Pos*, then re-runs the same return regressions as conducted in the above subsections. Here, we observe that the significance of *PoliticalPos* is now much reduced (reversed) by *Pos* on day [0] ([-1]). In Panel B, *Pos_YZZ* goes one step further: it reverses the effect of *PoliticalPos* on both days [-1, 0]. The reversal or subsumption is similar across all other days. That *Pos_YZZ* reverses the effect of *PoliticalPos* is perhaps not surprising, as the YZZ dictionary contains many of the political words within their identified positive words, *Pos_YZZ*. These results suggest that, as evidenced in our prior return association tests, our list of political words is "redundant" for return predictions, in that they do not provide incremental explanatory power to our more precisely defined (relative to alternative dictionaries) positive words. Overall, our evidence points to the non-uniqueness of political words, justifying our separation of those words into a separate list, as well as indicating that our Chinese dictionary is more precise than alternatives explored in this paper.

[Insert Table A6 about here.]

References

Baker, M., Litov, L., Wachter, J.A., and Wurgler, J. (2010) Can mutual fund managers pick stocks? Evidence from their trades prior to earnings announcements, Journal of Financial and Quantitative Analysis 45, 1111-1131.

Bergstra, J. and Bengio, Y. (2012) Random search for hyper-parameter optimization, Journal of Machine Learning Research 13, 281-305.

Bushee, B.J., Core, J.E., Guay, W. and Hamm, S.J.W. (2010) The role of the business press as an information intermediary, Journal of Accounting Research 48, 1–19.

Cortes, C. and Vapnik, V. (1995) Support-vector networks, Machine Learning 20, 273-297.

Huang, A.G., Tan, H. and Wermers, R. (2020) Institutional trading around corporate news: Evidence from textual analysis, The Review of Financial Studies 33, 4627–4675.

Kohavi, R. (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1137-1145.

Kothari, S.P., Shu, S. and Wysocki, P.D. (2009) Do managers withhold bad news? Journal of Accounting Research 47, 241–276.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013) Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems, 3111–3119.

Tetlock, P.C., Saar-Tsechansky, M. and Macskassy, S. (2008) More than words: Quantifying language to measure firms' fundamentals, Journal of Finance 63, 1437–1467.

Vapnik, V. (2013) The Nature of Statistical Learning Theory, Springer Science & Business Media, Berlin.

Yao, J., Feng, X., Wang, Z., Ji, R. and Zhang, W. (2021) Tone, sentiment and textual analysis: The construction of Chinese sentiment dictionary in finance (translated from mandarin), Journal of Management Sciences in China 24, 26-46.

You, J., Zhang, B. and Zhang, L. (2018) Who captures the power of the pen? The Review of Financial Studies 31, 43–96.

Table A1 Tokenization Comparison of Three Segmenters

This table shows the comparison of tokenization result of Jieba, NLPIR, and Stanford Word segmenter, where the benchmark is human reading (“Human”). The four sentences are related, respectively, to stock market commentary, share repurchase, real estate and firm legal issue.

Original Sentence	Segmenter	Segmentation	# of inconsistencies with human reading
从龙虎榜席位来看，在上市后的连续涨停板过程中，多个机构专用席位主要以卖出为主，买方多是游资。	Human	从/龙虎榜/席位/来看/，/在/上市/后/的/连续/涨停板/过程/中/，/多个/机构专用席位/主要/以/卖出/为主/，/买方/多是/游资/。	/
	Jieba	从/龙虎榜/席位/来看/，/在/上市/后/的/连续/涨停板/过程/中/，/多个/机构专用席位/主要/以/卖出/为主/，/买方/多/是/游资/。	1
	NLPIR	从/龙虎榜/席位/来看/，/在/上市/后/的/连续涨停/板/过程/中/，/多/个/机构专用席位/主要/以/卖/出/为主/，/买方/多/是/游资/。	3
	Stanford	从/龙虎/榜/席位/来/看/，/在/上市/后/的/连续/涨/停板/过程/中/，/多/个/机构/专用/席位/主要/以/卖出/为/主/，/买方/多/是/游资/。	4
以公司提取的购股资金和激励对象自筹配比的等额资金，从二级市场上回购自家公司股份授予激励对象。	Human	以/公司/提取/的/购股/资金/和/激励对象/自筹/配比/的/等额/资金/，/从/二级市场/上/回购/自家/公司股份/授予/激励对象/。	/
	Jieba	以/公司/提取/的/购股/资金/和/激励/对象/自筹/配比/的/等额/资金/，/从/二级市场/上/回购/自家/公司/股份/授予/激励/对象/。	3
	NLPIR	以/公司/提取/的/购/股/资金/和/激励/对象/自筹/配/比/的/等/额/资金/，/从/二级市场/上/回购/自家/公司/股份/授予/激励/对象/。	6
	Stanford	以/公司/提取/的/购股/资金/和/激励/对象/自/筹配/比/的/等额/资金/，/从/二/级/市场/上/回购/自家/公司/股份/授予/激励/对象/。	5
下行的二手房市场遭遇疫情“黑天鹅”，打折让利是最简单粗暴的抢夺市场的手段，如今的变化，似乎也预示着市场“暖冬”的苗头。	Human	下行/的/二手房/市场/遭遇/疫情/“/黑天鹅/”/，打折/让利/是/最/简单粗暴/的/抢夺/市场/的/手段/，/如今/的/变化/，/似乎/也/预示着/市场/“/暖冬/”/的/苗头/。	/
	Jieba	下行/的/二手房/市场/遭遇/疫情/“/黑天鹅/”/，/打折/让利/是/最/简单/粗暴/的/抢夺/市场/的/手段/，/如今/的/变化/，/似乎/也/预示着/市场/“/暖冬/”/的/苗头/。	1
	NLPIR	下/行/的/二手/房/市场/遭遇/疫情/“/黑/天鹅/”/，/打折/让利/是/最/简单/粗暴/的/抢夺/市场/的/手段/，/如今/的/变化/，/似乎/也/预示着/市场/“/暖冬/”/的/苗头/。	4
	Stanford	下行/的/二/手房/市场/遭遇/疫情/“/黑天/鹅/”/，/打折/让利/是/最/简单/粗暴/的/抢夺/市场/的/手段/，/如今/的/变化/，/似乎/也/预示/着/市场/“/暖冬/”/的/苗头/。	3
燕京啤酒公告，公司董事长、总经理赵晓东因涉嫌职务违法，被有关部门立案调查并采取留置措施，不能正常履职。	Human	燕京啤酒/公告/，/公司/董事长/、/总经理/赵晓东/因/涉嫌/职务违法/，/被/有关部门/立案调查/并/采取/留置措施/，/不能/正常/履职/。	/
	Jieba	燕京啤酒/公告/，/公司/董事长/、/总经理/赵晓东/因涉嫌/职务/违法/，/被/有关/部门/立案/调查/并/采取/留置/措施/，/不能/正常/履职/。	5
	NLPIR	燕京啤酒/公告/，/公司/董事长/、/总经理/赵晓东/因/涉嫌/职务/违法/，/被/有关/部门/立案/调查/并/采取/留/置/措施/，/不能/正常/履/职/。	5
	Stanford	燕京/啤酒/公告/，/公司/董事/长/、/总/经理/赵/晓东/因/涉嫌/职务/违法/，/被/有关/部门/立案/调查/并/采取/留置/措施/，/不能/正常/履职/。	8

Table A2   List of News Sources

This table lists 64 news sources that each supply at least 1,000 articles for full sample. We define a media outlet to be a state media if its ultimate control right belongs to the central or a provincial government.

Agency English name	Agency Chinese name	Full sample % (in 3,078,015 articles)	Firm-specific news % (in 682,945 articles)	State media
China Securities Journal	中国证券报	14.96	25.66	Yes
Securities Times	证券时报	22.40	24.00	Yes
Sina News	新浪新闻	24.17	12.38	No
Panorama Network	全景网络	4.75	6.31	Yes
Securities Daily	证券日报	5.28	4.30	Yes
Shanghai Securities News	上海证券报	2.71	3.00	Yes
National Business Daily	每日经济新闻	2.01	2.34	Yes
First Financial Daily	第一财经日报	2.32	2.11	No
21st Century Business Herald	21 世纪经济报道	2.18	2.03	No
Jiemian News	界面新闻	1.03	1.47	Yes
Gelonghui	格隆汇	0.46	1.33	No
The Beijing News	新京报	1.10	1.31	Yes
Zhitong Finance	智通财经网	0.47	1.01	No
Beijing Business Today	北京商报	0.43	0.88	Yes
China Business Journal	中国经营报	0.44	0.84	No
The Paper	澎湃新闻	0.38	0.82	Yes
China News Service	中国新闻网	0.70	0.78	Yes
Economic Daily	经济日报	0.50	0.73	Yes
EastDay	东方网	0.40	0.68	Yes
Xinhua News Agency Web	新华网	0.69	0.59	Yes
Great Wisdom Information Technology	大智慧	0.62	0.55	No
The Economic Observer	经济观察报	0.32	0.51	No
China Times	华夏时报	0.52	0.50	Yes
The State Council Information Office	中国网	0.68	0.43	Yes
International Finance News	国际金融报	0.24	0.43	Yes
Capital Week	证券市场周刊	0.49	0.38	Yes
Finet Group Limited	财华网	0.17	0.38	No
Investment Express	投资快报	0.40	0.37	Yes
The Time Weekly	时代周报	0.22	0.25	Yes
Chinese Securities Journal	大众证券报	0.22	0.22	Yes
GuangZhou Daily	广州日报	0.23	0.22	Yes
Economic Information Daily	经济参考报	0.77	0.22	Yes
Yangcheng Evening News	羊城晚报	0.27	0.20	Yes
Caijing Magazine	财经网	0.20	0.20	Yes
CHINAFUND	中国基金报	0.16	0.20	Yes
Beijing Times	京华时报	0.33	0.19	Yes
People's Daily	人民日报	0.35	0.19	Yes
ChangJiang Times	长江商报	0.18	0.17	Yes
Investor Journal	投资者报	0.13	0.17	No

Bull and Bear Trading Room	牛熊交易室	0.05	0.16	No
Moneyweek	理财周报	0.20	0.14	Yes
Hexun News	和讯网	0.60	0.14	No
Information Times	信息时报	0.16	0.14	Yes
China Investment Network	投资时报	0.12	0.12	Yes
Civil Aviation Resource Net of China	民航资源网	0.07	0.11	No
Southern Metropolis Daily	南方都市报	0.22	0.10	Yes
Global Times	环球时报	0.07	0.09	Yes
Cailian Press	财联社	0.05	0.08	Yes
Jinling Evening News	金陵晚报	0.08	0.08	Yes
Wallstreetcn	华尔街见闻	0.09	0.08	No
Shanghai Morning Post	新闻晨报	0.06	0.07	Yes
Beijing Youth Daily	北京青年报	0.07	0.06	Yes
China Economic Weekly	中国经济周刊	0.16	0.06	Yes
Chongqing Economic Times	重庆商报	0.06	0.05	Yes
China Economic Herald	中国经济导报	0.06	0.04	Yes
Legal Weekly	法治周末	0.07	0.03	Yes
The General Office of the State Council	中国政府网	0.07	0.03	Yes
Qianjiang Evening News	钱江晚报	0.08	0.03	Yes
The Mirror	法制晚报	0.10	0.02	Yes
China Central Television	央视网	0.05	0.02	Yes
China National Radio	中国广播网	0.09	0.02	Yes
China Finance Information	中财网	0.08	0.02	No
National Bureau of Statistics of China	国家统计局网站	0.08	0.00	Yes
Leju Finance	乐居财经	0.05	0.00	No
Others	Others	3.33	3.77	/

Table A3 Days [0, 1] Returns and Alphas of Portfolios Sorted on Sentiment

We sort the sample into decile portfolios on the ranked value of sentiment, and calculate the equal-weighted returns and alphas per the Capital Asset Pricing Model on the decile portfolios. “D1–D10” is the difference between decile 1 and decile 10. *T*-statistics are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Decile	1(s)	2	3	4	5	6	7	8	9	10(L)	D1–D10
Panel A: *Neg_net* returns and alphas											
Day [0] return	1.63	0.93	0.71	0.44	0.23	0.12	0.03	-0.07	-0.21	-0.48	2.11*** (75.35)
Day [0] CAPM alpha	1.48***	0.82***	0.62***	0.36***	0.18***	0.07***	-0.01	-0.09***	-0.23***	-0.51***	1.99*** (80.66)
	(80.14)	(48.56)	(38.68)	(23.64)	(12.14)	(5.05)	(-0.70)	(-6.60)	(-15.90)	(-31.09)	
Day [1] return	0.28	0.19	0.15	0.08	0.05	0.07	0.08	0.05	0.03	0.04	0.25*** (9.77)
Day [1] CAPM alpha	0.22***	0.10***	0.07***	0.02	-0.01	-0.02	-0.02	-0.05***	-0.07***	-0.12***	0.34*** (15.56)
	(14.01)	(6.50)	(4.67)	(1.33)	(-0.87)	(-1.44)	(-1.59)	(-3.69)	(-4.88)	(-7.85)	
Panel B: alphas for *Neg, Pos,* and *PoliticalPos*											
Day [0] CAPM alpha on *Neg*	0.95	0.58	0.53	0.41	0.27	0.22	0.14	0.06	-0.08	-0.43	1.38*** (59.64)
Day [0] CAPM alpha on *Pos*	-0.12	-0.10	-0.10	-0.08	-0.01	0.14	0.29	0.50	0.74	1.44	-1.56*** (-66.81)
Day [0] CAPM alpha on *PoliticalPos*	0.16	0.01	0.11	0.13	0.17	0.27	0.26	0.37	0.49	0.72	-0.56*** (-24.08)
Day [1] CAPM alpha on *Neg*	0.10	0.05	0.06	0.04	0.01	0.02	0.01	0.00	-0.07	-0.10	0.20*** (9.78)
Day [1] CAPM alpha on *Pos*	-0.05	-0.04	-0.05	-0.03	-0.05	-0.04	0.02	0.05	0.10	0.20	-0.25*** (-12.21)
Day [1] CAPM alpha on *PoliticalPos*	-0.05	-0.06	-0.03	-0.00	-0.01	0.04	0.05	0.04	0.06	0.08	-0.14** (-6.67)

Table A4 Robustness Checks of Return Regressions

This table reports the regression results of Table VII in the main text for various samples. Panel A uses news clustered with a minimum of three days of stopping interval between two news clusters per Huang, Tan, and Wermers (2020). There are 199,370 observations for day [0] at Panel A. Panel B (C) uses the sample of press-initiated (firm-initiated) news only, and has 243,532 (225,452) observations for day [0]. Panel D calculates the sentiment measures from the news headlines only; Panel E removes all intra-day news from Table VII; and Panel F removes all news occurring [-3, 3] trading days around an earnings announcement from Table VII. Each line represents a return regression of Table VII; the control variable results are omitted for brevity. *T*-statistics are two-way cluster-adjusted at the firm and news-date levels, and are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Return regressions using clustered news									
	Industry- and size-adjusted return over day(s)								
	[-10, -6]	[-5, -3]	[-2]	[-1]	[0]	[1]	[2]	[3, 5]	[6, 10]
Neg_net	-1.098***	-1.066***	-1.127***	-5.407***	-12.021***	-0.061	0.851***	0.638***	0.173**
	(-14.79)	(-11.89)	(-7.10)	(-27.61)	(-52.54)	(-0.40)	(5.62)	(7.85)	(2.57)
Panel B: Press-initiated news only									
Neg_net	-1.236***	-1.939***	-2.955***	-8.432***	-12.509***	-1.783***	-0.016	0.551***	0.212***
	(-16.30)	(-20.29)	(-20.78)	(-42.55)	(-43.74)	(-12.51)	(-0.12)	(7.12)	(3.59)
Panel C: Firm-initiated news only									
Neg_net	-0.645***	-0.497***	-0.350*	-3.401***	-9.513***	-1.142***	0.191	0.502***	0.051
	(-6.62)	(-4.28)	(-1.84)	(-12.78)	(-36.82)	(-6.51)	(1.02)	(5.37)	(0.78)
Panel D: Using sentiment in news headlines only									
Neg_net	-0.326***	-0.449***	-0.530***	-2.136***	-4.266***	-0.655***	-0.045	0.149***	0.065***
	(-12.87)	(-14.57)	(-9.96)	(-30.01)	(-46.17)	(-14.12)	(-1.04)	(6.05)	(3.46)
Panel E: Removing all intra-day news									
Neg_net	-1.271***	-1.724***	-2.231***	-7.352***	-6.005***	-0.491***	0.223*	0.470***	0.078
	(-17.20)	(-18.38)	(-15.16)	(-34.34)	(-34.92)	(-3.83)	(1.72)	(6.92)	(1.47)
Panel F: Excluding news [-3, 3] days around earnings announcements									
Neg_net	-1.216***	-1.570***	-2.098***	-6.763***	-11.459***	-1.429***	0.081	0.571***	0.148***
	(-17.27)	(-17.72)	(-15.35)	(-35.96)	(-49.68)	(-11.06)	(0.63)	(8.28)	(2.85)

Table A5 Return Regressions of Sentiments from Alternative Dictionaries

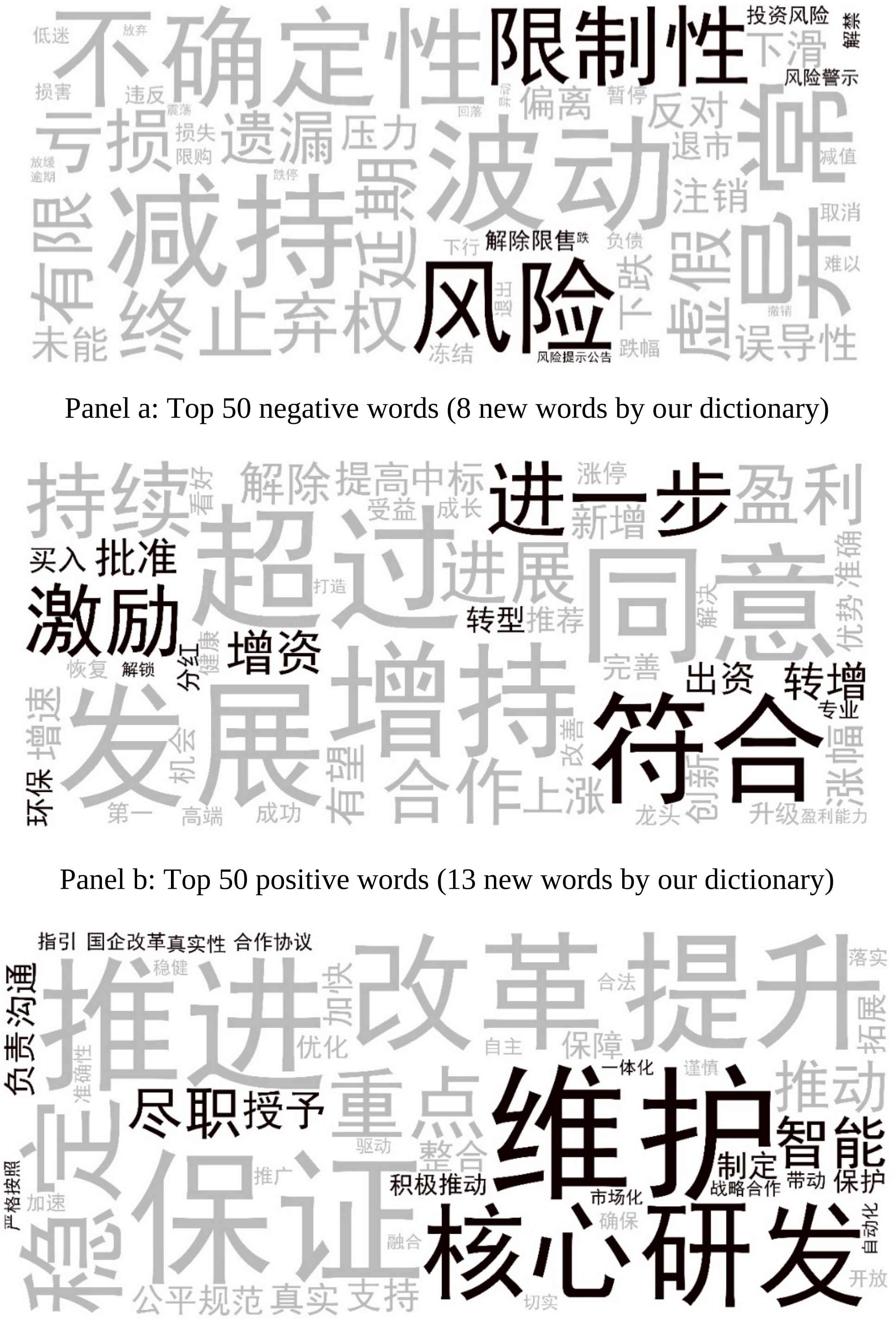
This table reports the regression results of Table VII of the main text for sentiment measures from LM, YZZ, and Generic dictionaries. Each cell represents a regression. The control variable results are omitted for brevity. *T*-statistics are two-way cluster-adjusted at the firm and news-date levels, and are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Industry- and size-adjusted return over day(s)								
	[-10, -6]	[-5, -3]	[-2]	[-1]	[0]	[1]	[2]	[3, 5]	[6, 10]
Neg_net_LM	-1.098***	-1.158***	-0.624***	-3.821***	-8.470***	-0.914***	0.354*	0.322***	0.032
	(-10.23)	(-8.85)	(-2.88)	(-13.28)	(-35.66)	(-5.09)	(1.69)	(3.42)	(0.43)
Neg_LM	-1.647***	-1.748***	-0.992***	-4.978***	-12.540***	-1.278***	0.466**	0.503***	0.203**
	(-12.51)	(-11.31)	(-3.55)	(-14.08)	(-38.03)	(-5.67)	(2.18)	(4.10)	(2.04)
Pos_LM	0.158	0.149	-0.038	1.937***	1.503***	0.294	-0.122	-0.012	0.281**
	(0.96)	(0.70)	(-0.12)	(5.27)	(3.92)	(1.03)	(-0.31)	(-0.08)	(2.27)
Neg_net_YZZ	-1.000***	-1.320***	-1.914***	-5.784***	-9.026***	-1.208***	0.153	0.441***	0.140***
	(-17.00)	(-17.56)	(-17.41)	(-37.78)	(-48.38)	(-10.92)	(1.30)	(7.59)	(3.10)
Neg_YZZ	-1.002***	-0.992***	-1.456***	-5.451***	-12.237***	-1.362***	0.342**	0.615***	0.168**
	(-9.28)	(-7.98)	(-6.87)	(-22.15)	(-41.45)	(-7.25)	(1.98)	(6.10)	(2.03)
Pos_YZZ	1.265***	1.827***	2.638***	7.445***	9.810***	1.448***	-0.103	-0.469***	-0.162***
	(16.23)	(17.74)	(18.86)	(36.01)	(40.51)	(9.87)	(-0.61)	(-6.25)	(-2.87)
Neg_net_generic	-0.358	-0.342	-0.823	-4.656***	-6.866***	-2.498***	0.706	0.287	-0.277
	(-1.07)	(-0.86)	(-1.40)	(-7.06)	(-8.60)	(-4.34)	(0.92)	(0.88)	(-1.13)
Neg_generic	-10.905***	-13.277***	-17.031***	-28.834***	-43.454***	-2.645**	0.679	2.555***	1.299**
	(-15.41)	(-15.12)	(-12.25)	(-19.49)	(-25.13)	(-2.07)	(0.52)	(3.40)	(2.24)
Pos_generic	-1.376***	-1.848***	-1.884***	-0.452	-0.758	1.723***	-0.515	0.167	0.441*
	(-4.39)	(-5.03)	(-3.37)	(-0.75)	(-1.05)	(3.26)	(-0.71)	(0.56)	(1.90)

Table A6 Non-Uniqueness of *PoliticalPos* in Return Regressions

We run the return regression of Table VII of *PoliticalPos* with the additional control of *Pos* or *Pos_YZZ*. The control variable results are omitted for brevity. *T*-statistics are two-way cluster-adjusted at the firm and news-date levels, and are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Industry- and size-adjusted return over day(s)								
	[-10, -6]	[-5, -3]	[-2]	[-1]	[0]	[1]	[2]	[3, 5]	[6, 10]
Panel A: *PoliticalPos* controlled for *Pos*									
PoliticalPos	-0.872***	-1.243***	-0.936***	-1.042***	0.920***	0.198	-0.091	0.028	0.146*
	(-7.13)	(-8.40)	(-4.56)	(-4.30)	(3.48)	(1.00)	(-0.33)	(0.25)	(1.79)
Pos	1.921***	2.753***	3.622***	9.122***	12.525***	1.677***	-0.024	-0.707***	-0.305***
	(21.60)	(23.98)	(21.80)	(40.84)	(44.22)	(9.93)	(-0.16)	(-8.26)	(-4.75)
Panel B: *PoliticalPos* controlled for *Pos_YZZ*									
PoliticalPos	-1.562***	-2.246***	-2.209***	-4.615***	-2.281***	-0.366	-0.023	0.140	0.211**
	(-11.31)	(-13.37)	(-9.18)	(-16.09)	(-7.30)	(-1.58)	(-0.08)	(1.10)	(2.28)
Pos_YZZ	1.875***	2.705***	3.503***	9.252***	10.703***	1.592***	-0.094	-0.524***	-0.244***
	(20.41)	(22.57)	(20.20)	(38.12)	(38.69)	(8.78)	(-0.58)	(-5.69)	(-3.61)



Panel a: Top 50 negative words (8 new words by our dictionary)

Panel b: Top 50 positive words (13 new words by our dictionary)

Panel c: Top 50 political words (21 new words by our dictionary)

Figure A1: Word clouds of the top 50 positive, negative and political words. Words appearing in the three alternative dictionaries (LM, YZZ, and Generic) are in gray font, and the original words in our dictionary are in bold font.