Textual analysis for China’s financial markets: a review and discussion

Alan Huang
University of Waterloo, Waterloo, Canada

Wenfeng Wu
Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai, China, and

Tong Yu
Department of Finance, Lindner College of Business, University of Cincinnati, Cincinnati, Ohio, USA

Abstract
Purpose – This is a literature survey paper. The purpose of this paper is to focus on the latest developments in textual analysis on China’s financial markets, highlighting its differences from existing works in the US markets.

Design/methodology/approach – The authors review the literature and carry out an experiment of sentiment analysis based on a small sample of Chinese news articles.

Findings – Based on the experiment of sentiment analysis, there is limited evidence on the association between sentiment and other contemporaneous or future returns.

Originality/value – The supply of financial textual information has grown exponentially in the past decades. Technological advancements in recent years make the programming-based analysis an effective tool to digest such information. The authors highlight the use of credible textual information and discuss directions of research in this important field.

Keywords Textual analysis, Natural language processing, Chinese financial market, Information content of news

Paper type Literature review

1. Introduction
The past decade witnesses an outburst of computational linguistics applications. While voice-based linguistics applications such as Apple’s Siri, Amazon’s Alexa and Google Assistant have become household names, many more applications and methodologies are developed in the textual processing. This field of computational linguistics which involves programming computers to process and analyze verbal and written languages is often called natural language processing (hereafter NLP).

Financial sector supplies a large amount of textual information. This textual information includes macro policy announcements, press news, firm announcements, regulatory filings, analyst reports and social media posts, to name just a few. Take the news releases of an individual stock listed in Shanghai Stock Exchange, Shanghai Pudong Development Bank, as an example. On the single day of July 31, 2019 alone, finance.sina.com.cn displays six firm-specific news releases, and 83 industry-related news articles on the firm. Multiplying these numbers by 3,650 – the number of listed firms in Shanghai Stock Exchange and Shenzhen Stock Exchange as of July 2019 – would yield an amount of articles (news stories

The authors thank the excellent research assistance provided by Joey Bao. Part of the work is completed while Huang was visiting Shanghai Jiao Tong University. The authors acknowledge financial support from the National Natural Science Foundation of China (Nos 71790592; 71850010) and Shanghai Institute of International Finance and Economics.
Textual analysis arguably facilitates more intelligent decision making. Take the asset management industry, for example. As of 2017, China’s asset management industry has an assets-under-management size of RMB 53.6 trillion, about two-thirds of its GDP\(^1\). Quantitative decision making is one leading trend of asset management; yet it is largely constrained to structured data so far. Textual analysis turns large amounts of unstructured data that the industry relies on everyday to structured output, thus enriching the quantitative decision-making process. Efficiently consuming embedded information of financial texts and possibly enhancing the trading strategies via text mining may significantly improve the risk-adjusted returns of the asset management industry. A simple application, for instance, can be that portfolio managers, who manually follow a wide range of financial news of their portfolio companies, can turn to a summary of news sentiment and topics via the output of text mining.

In this review paper, we provide a summary of the research status of financial textual analysis research for China’s financial markets. Although in practice we tend to view financial textual information as important as quantitative data, large scale processing of financial textual information has received much less attention, partly due to the inherent difficulty in dealing with textual data. We provide a review of existing tools and current research papers, and showcase a simple application of sentiment analysis in China’s A-share market. A caveat is that this review paper does not intend to include each and every NLP paper related to Chinese financial markets; rather, we aim to provide, with the hope to stimulate NLP research in Chinese financial texts, the review of the general research framework and include, to the best of our knowledge, the representative papers. We also emphasize due to the potential to tie financial textual research to monetary decision making, researchers should make efforts to ensure integrity in research designs and keep in mind the accuracy of their research output.

2. Financial textual mining: an overview
Different from quantitative information, textual data are unstructured in nature, making its analysis difficult to conduct (Chan and Chong, 2017). To extract meaningful information from textual data, one must decipher a narrative’s word order and meaning. Within the realm of accounting and finance, textual analysis methodologies have significantly evolved over the past decade, largely attributable to the improvement in technological capabilities (e.g. Loughran and McDonald, 2016). Researchers extract quantitative metrics of financial text documents from a number of angels, most notably file size (how large is the file), readability (how difficult it is to comprehend the text) and sentiment (how positive or negative the tone of the document is).

Loughran and McDonald (2016) provide a comprehensive survey of text mining in the academic accounting and finance literature in the US market. By and large, the current literature examines the following aspects of texts: readability score, bags of words that have different sentiment (e.g. negative, positive, strong and weak words), collapsing a document to term-document matrix consisting of rows of words and columns of word counts, comparing how similar two documents are based on term-document matrix (e.g. cosine similarity) and thematic structure in documents that leads to topic modeling (e.g. what topic the document is about). Notably, Loughran and McDonald (2011) develop financial keyword lists (dictionaries such as negative and positive sentiment words), and show that their lists of financial keywords are significantly better at capturing the content in financial texts than the general purpose, commonly used Harvard IV-4 word dictionary. Loughran and McDonald’s (2011) word lists have since been widely used in the accounting and finance literature[2].
Extant research using the data in the US market has shown that file size, readability and sentiment of earnings reports, annual reports, periodic material event filings and corporate news are related to various aspects of firms, in particular, to stock returns. For example, Tetlock et al. (2008) examine the “tone” of corporate news (by counting negative and positive words), and find that it predicts the cross-section of stock returns. Studies also link news content to return momentum (e.g. Chan, 2003) and return reversals (e.g. Tetlock, 2007, 2011). In other forms of texts, for instance, Kim et al. (2019) find that less readable 10-K reports are associated with higher stock price crash risk, and Zhao (2016) finds that higher Form 8-K (material event) filing frequency is associated with lower future returns. Research also finds that management gives preferred access to favored analysts in conference calls (e.g. Mayew et al., 2019). Mayew and Venkatachalam (2012) use vocal emotion analysis software to analyze call audio files and find that audio tone displayed by managers is informative about the firm’s financial future such as return on assets.

Echoing the academic research, on the application side, financial textual analysis in the USA yields a number of commercial products, most of which focus on one specific type of text source. On SEC filings, data provider Wharton Research Data Services produces a suite of sentiment and readability scores from firms’ historical SEC filings archives; and the start-up firm SeekEdgar provides similar yet up-to-date sentiment and readability scores of SEC filings. On the corporate news front, Thomson Reuters, a main-stream news provider, and Ravenpack, a start-up business intelligence firm, both provide sentiment scores of US firm and macro news for financial clients. Thomson Reuters’ news product, called Thomson Reuters News Analytics, contains news sentiment scores for a broad range of equities (with a strong focus on the US market), and is marketed as “the only service of its kind of the market” on their flyer. By and large, the financial-text-mining commercial products in the US market focus on a specific type of text (e.g. SEC filings or corporate news) and provide metrics that are proven in academic studies (e.g. sentiment and readability scores).

3. Current status of financial textual mining in the Chinese markets

One lesson from the US market is that financial texts are significantly different from the general texts and require domain-specific lexicons (i.e. dictionaries) for NLP. For example, while the word “fine” in everyday English generally denotes a neutral sentiment, it mostly implies a negative sentiment in financial text (e.g. a firm is “fined” by SEC). As well, every single financial market is regulated differently. Regulation impacts information flow and information conveying; and hence the forms and writing of financial texts are also relatively market specific. NLP should take these elements into consideration.

In both mainland China and Hong Kong, not only is the financial sector important to their respective economy, each of their market capitalizations ranks among the top 10 capital markets worldwide. Given the significance of their financial markets worldwide, and that they share a common official language (Chinese), textual analysis of financial information is important. Financial textual analysis for Chinese markets (Hong Kong included), however, is sparse.

Let us discuss, first, Hong Kong, which has a more developed financial market than mainland China. We believe that a number of elements lead to the underdevelopment of financial textual research and applications in Hong Kong. Listed firms in Hong Kong file with Hong Kong Exchanges and Clearing Limited’s dedicated filing system, called HKExnews, in PDF format in both English and traditional Chinese languages. Parsing PDF files can pose a challenge as they often contain images and tables, and often have non-standard textual layouts. As well, the multi-language environment in Hong Kong – where news and filings appear in three languages (English, traditional Chinese and simplified Chinese) – amplifies the difficulty in dealing with textual data. On the academic side, there are limited studies that tackle Hong Kong’s financial text; and the ones that we are able to identify use either social media data or rely on a very limited number of firms and news[3]. On the business side, top
banks and data aggregators tend to treat Hong Kong as a sales office rather than an R&D hub. These reasons combined result in less cutting-edge financial research using the Hong Kong market data; and particularly so for emerging areas such as NLP. For example, in a recent conference with the theme “AI and Sentiment Analysis in Finance” organized by UNICOM Seminars Ltd that took place in March, 2018 in Hong Kong, the talks were mostly about general techniques and applications in the US market; actually, none of the talks was about the Hong Kong market (http://conferences.unicom.co.uk/sentiment-analysis-hongkong-2018).

Likewise, financial textual analysis for the Mainland China market is at its infancy. One reason is that the Chinese language is distinctly different from the English language. For example, the Chinese language expresses meanings by the combination of individual characters, making even the simple task of separating the English equivalent of “words” – which is essentially the first step in textual analysis – difficult. We, however, see a burgeoning of financial textual research in Mainland China in recent years.

Earlier work of Chinese financial textual analysis starts with manually analyzing text contents for specific topics. For example, in investigating the relationship between the social responsibility disclosure index and firm characteristics, Shen (2007) manually constructs social responsibility disclosure index by reading the corresponding parts of social responsibility in 655 annual reports of Chinese listed companies. He et al. (2016) manually analyze the content of 228 Chinese listed firms’ Weibo (micro-blogs) and examine the role of social media in information disclosure. You et al. (2018) find a regional bias in Chinese media report news by manually judging the news text. Being subjective aside, manual examining textual analysis is notwithstanding time-consuming, constraining the amount of texts that can be assimilated.

Subsequent Chinese textual analysis of the literature focuses on the sentiment of texts by counting words with different tones following Loughran and McDonald (2011). Studies analyze how the tone of texts from IPO prospectuses, annual report, news and social media affects variables such as returns, earning and insider trading, etc. (Xie and Lin, 2015; Wang and Wu, 2015; Lin and Xie, 2016, 2017; Yan et al., 2018; Zhu and Xu, 2018). As to texts from IPO prospectuses, Yan et al. (2018) examine whether and how the tone of IPO prospectuses affects stock return and risk after IPOs. They find that uncertain or negative tones in IPO prospectuses are significantly associated with IPO initial returns and the post-IPO return volatility. Such negative or uncertain tones in the prospectuses also lower firms’ long-term returns. Zou et al. (2019), however, argue that it is media coverage during an IPO which significantly negatively associated with IPO underpricing as well as post-IPO volatility, regardless of whether the tone is positive or negative. Wang and Wu (2015) investigate the effects of the sentiments of news around the IPO date of 728 Chinese listed firms on IPO underpricing. They find that the media tone is negatively and significantly related to IPO underpricing, over-subscription and underwriter proceeds.

As to texts from annual reports, Zhu and Xu (2018) find that managers’ tone of optimism (pessimism) in annual reports is associated with upward (downward) earning management. Another text source is earning communication conferences of Chinese listed firms, which Xie and Lin (2015) and Lin and Xie (2016, 2017) use to measure management tone by sentiment words counting. They find that management tone is significantly associated with the next year’s earning (Xie and Lin, 2015), post-call cumulative abnormal returns (Lin and Xie, 2016) and analysts’ recommendation revisions (Lin and Xie, 2017). Zhang and Huang (2018) find that sentiment index constructed by the tone of Fintech-related news articles is negatively associated with online lending default rate. The association is significantly not only for matured loans, but also for new loans.

Word counting of sentiment relies on “bag of words” and a well-defined financial word lexicon in Chinese – such as dictionaries for sentiment words and for uncertain words, etc. At the time of writing, there is not a well-accepted Chinese financial sentiment word dictionary. Perhaps to make matters worse, even general-purpose sentiment words in
Chinese are not well defined. There exist a few general-purpose sentiment word dictionaries, such as those from Dalian University of Technology and HowNet; however, there is little conformity across these different dictionaries (see the elaboration in the next section). For financial sentiment words, a number of studies use the Chinese translation of Loughran and McDonald (2011) lexicons (Wang and Wu, 2015; Xie and Lin, 2015; Lin and Xie, 2016; Zeng et al., 2018; Li et al., 2019). The power of such translated list of words is dubious, as the language expressions are significantly distinct across Chinese and English. In an effort to ameliorate this problem, Li et al. (2019) augment the Chinese translation of Loughran and McDonald (2011) sentiment word list with manually identified sentiment words from message boards and find it fits the slang language nature of message boards better.

More recent approaches in machine learning are utilized to construct lexicons. Using Word2Vec method by Mikolov et al. (2013), Wang and Huang (2018) construct tone lexicon of Chinese words in peer-to-peer (P2P) lending to analyze how the news tones affect trading volume of P2P in China. More generally, Yao et al. (2019) use deep learning long short-term memory model to construct sentiment lexicon of Chinese words in finance fields based on annual reports of Chinese listed firms and social media, respectively. Essentially, Yao et al. (2019) infer the significance of potential word sentiment from the stock returns in the three-day window post annual report announcements. On potential drawback of this approach is that a majority of listed firms either “pre-announce” or release management forecasts of the key financial metrics such as revenue and net income some time before the formal annual report releases – such pre-announcement leads to investor expectation, rendering formal annual report releases less return-informative.

In addition to word counting, studies also extend to the sentence- or message-level sentiment using machine learning classification methods, such as Naïve Bayesian, support vector machine (SVM) and K-nearest neighbors classification (KNN) to analyze the text contents, Jin et al. (2013) crawl nearly 6m messages posted on GUBA, the largest stock message board in China, and use KNN method to classify each message into “optimistic,” “pessimistic” and “neutral.” They find that optimistic opinion is associated with higher stock returns and that diversified opinion is associated with higher trading volume. Using the KNN method, Shen (2018) finds that messages posted on the social media by star analysts have a significant effect on the stock price. Both studies conclude that KNN is better than Naïve Bayesian in uncovering the text content. Following Hanley and Hoberg (2010), Meng et al. (2017) construct words vectors in the “Management Discussion and Analyses” section in annual reports of Chinese listed firms to proxy for text information content. They decompose such information content into market-related, industry-related and idiosyncratic components, and find that firms with the higher idiosyncratic information component have lower future crash risk. Hao and Su (2014) use a similar method to analyze IPO prospectuses. Perhaps the largest sample in this area is Li et al. (2019). These authors use SVM and convolutional neural network to infer message-level sentiment of 60m social media posts on Eastmoney (a leading online investor forum). They manually classify a small sample of message posts into different sentiment category and use the manual classification as the training input to classify the entire sample of message posts.

On the application side, there exists the so-called “public opinion monitoring” by a number of firms, which basically counts the number of news articles and flags the article to be either positive or negative in sentiment. WIND financial terminal (the largest financial data aggregator in China, an equivalent of Bloomberg) represents one such usage.

4. An experimental application of sentiment analysis in the Chinese A-share market
In this section, we carry out an experiment of sentiment analysis with a small sample of Chinese corporate news. Our analysis is exploratory in nature – we aim to bring to the attention of readers issues such as data availability, dictionaries and possible usefulness of firm news.
4.1 Sample and dictionaries
We download 50 Chinese news articles each for the SSE 50 index firms. The SSE 50 index consists of 50 largest firms listed in the Shanghai Stock Exchange. To do so, we write a web scraper script to download firm-specific news articles from finance.sina.com.cn, a popular finance website that streams real-time news and stock information for each stock. We require that each news article have at least 100 words for it to be meaningfully parsed. We download articles inverse-chronologically, starting from April 1, 2019, and stop until we have 50 articles for each firm. Our sample thus has 2,500 news articles.

Figure 1 plots the date distribution of the news articles. We note that the majority – 80.5 percent to be exact – of the news articles appear in March and April 2019 (with 323 news articles, or 13 percent, appearing on April 1, 2019, the last news event date); however, the farthest article dates back to May 28, 2018. Therefore, there seems a sufficient dispersion of event dates for an event study. Nonetheless, the news date distribution in Figure 1 suggests there are a large number of firm-specific news articles for large firms such as those in SSE 50. Huang et al. (2019) show the complexity of dealing with successive news, and caution that inferences drawn from successive news need to be carefully interpreted. In particular, they advocate using the first news in a news series to pin down the starting event time. With this in mind, we nonetheless proceed with usual event study methodologies in this exploratory work.

In this experiment of sentiment analysis, our main job next is to calculate a sentiment score for each of the articles. The first step is to parse the news article by removing the stop words, which are commonly used words that do not carry “real” meanings (such as “me,” “you”), and then by tokenizing the text corpuses, which essentially breaks the text into words. We use the stop-word list from www.sogou.com, a Chinese search engine that contains lists of keywords in various areas. There are two often-used tokenization tool for
Chinese words, one is Python package Jieba, and the other is NLPIR-ICTCLAS by Chinese Academy of Sciences. We use Jieba for this exercise.

After tokenizing the text corpuses, the second step is to count for each news article the number of negative and positive word occurrences, in which we need to rely on a given dictionary (lexicon). Since there does not exist a commonly accepted Chinese financial word dictionary, we instead use more general sentiment word dictionaries. We choose three such sentiment word dictionaries, by, respectively, Dalian University of Technology (hereafter “Dalian”), China National Knowledge Infrastructure (hereafter “HowNet”) and National Taiwan University (hereafter “NTU”). These dictionaries are widely used in common settings. Concurrent to this review paper, Yao et al. (2019) create a Chinese financial word dictionary using a large sample of corporate annual reports and social media posts based on machine learning methods. Their list of words is not known to us; they, however, show that the negative sentiment ratio calculated from their list of word is significantly related to those calculated from Dalian and HowNet dictionaries.

Loughran and McDonald (2011) compile a list of positive and negative English financial words. We accordingly keep only words with positive and negative sentiments from the aforementioned dictionaries. Where appropriate, we drop the degree of emotion (i.e. valence) attached to each word and simply classify sentiment words into either positive or negative. It is worth noting that the three dictionaries have significant differences, as shown in Table I.

We make a number of observations from Table I. Dalian has the largest dictionary. Its total number of sentiment words is close to 22,000, and these words are about evenly split between positive and negative words. This is much larger than the sentiment words in the Loughran and McDonald (2011) dictionary, which has only 2,355 negative words and 354 positive words. Upon scanning the list of the words in these dictionaries, we note that many of the words seem to fit with general texts rather than financial texts. This point is similarly emphasized in Loughran and McDonald (2011) – that a general-purpose English word dictionary such as Harvard IV-4 is not suitable for financial texts. Recently, Yao et al. (2019) form a Chinese financial word dictionary of 1,633 negative words and 3,592 positive words based on about 20,000 Chinese firms’ annual reports, which is also much smaller than the number of the words in the three general-purpose dictionaries.

Another distinctive observation is that the three dictionaries overlap little. There are 43,335 unique words in all three dictionaries, but only 1,205 of them, or 2.78 percent, appear in all three dictionaries. We define these 1,205 as the “common” dictionary. The little agreement among the three widely used dictionaries suggests that even the general Chinese sentiment word list has no consensus. We believe that part of the reason is that Chinese words connote meaning by combinations of Chinese characters. One frequent effect of such combination is that two words are likely to convey the same sentiment as long as they contain the same core characters, even though the number of characters of the two words may be different. One way

<table>
<thead>
<tr>
<th></th>
<th>Dalian</th>
<th>HowNet</th>
<th>NTU</th>
<th>Total</th>
<th>Common</th>
<th>% Common</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of words</td>
<td>21,908</td>
<td>8,746</td>
<td>20,764</td>
<td>43,335</td>
<td>1,205</td>
<td>2.78</td>
</tr>
<tr>
<td>No. of positive words</td>
<td>11,230</td>
<td>4,484</td>
<td>9,509</td>
<td>21,194</td>
<td>639</td>
<td>3.02</td>
</tr>
<tr>
<td>No. of negative words</td>
<td>10,678</td>
<td>4,262</td>
<td>11,255</td>
<td>22,141</td>
<td>566</td>
<td>2.56</td>
</tr>
<tr>
<td>No. of words, 1 ≤ 2 syllables</td>
<td>7,485</td>
<td>5,388</td>
<td>9,302</td>
<td>17,120</td>
<td>883</td>
<td>5.16</td>
</tr>
<tr>
<td>No. of words, &gt; 2 syllables</td>
<td>14,423</td>
<td>3,358</td>
<td>11,462</td>
<td>26,215</td>
<td>322</td>
<td>1.23</td>
</tr>
</tbody>
</table>

**Notes:** Dalian, HowNet and NTU are three general Chinese language sentiment word dictionaries. The column “Total” refers to the total unique words, the column “Common” refers to words that appear in all three dictionaries and the column “% Common” is the column “Common” divided by the column “Total”.

**Table I.** Comparison of the sentiment word list across Dalian, HowNet and NTU dictionaries.
to ameliorate this problem is through careful tokenization of texts (i.e. separating sentences into appropriate words). Dealing with more homogeneous texts, such as financial texts, is helpful as well, as the sentiment words tend to be the same across documents.

The low degree of overlapping of dictionaries may also be ameliorated through examining shorter words. We hence separate the words into long and short ones, defined, respectively, as words with more than two syllables and otherwise. Table I shows that the total number of long words is somewhat larger than that of short words, and that the common ratio of short words across the three dictionaries is indeed much higher than that of long words. The common ratio of short words, however, still remains very low at 5.16 percent. Overall, the evidence in Table I shows that using the common dictionaries for financial textual analysis (or even for general sentiment analysis) is probably not ideal.

We use Dalian and HowNet dictionaries and the common dictionary to count sentiment words for each news article. Then in the next step we construct a measure of news tone, $\text{Neg}_\text{net}$, as the fraction of total negative word occurrences (including those in the headline and body of the news) net of total positive word occurrences in each news article (Huang et al., 2019), i.e.:

$$\text{Neg}_\text{net} = \frac{\text{No. of negative word occurrences} - \text{No. of positive word occurrences}}{\text{No. of total words in the news article}}.$$  

Since the literature typically emphasizes negative words (e.g. Tetlock et al., 2008), we also consider $\text{Neg}$, the ratio of negative word count to total number of words in the news article. Obviously, $\text{Neg}_\text{net}$ is bounded between minus one and plus one, and $\text{Neg}$ is bounded between zero and one. A higher ratio indicates that the news is more negative. Table II shows the distribution of the sentiment measures.

Table II shows that $\text{Neg}_\text{net}$ tends to be negative, or that news tends to be positive, consistent with the literature and the general perception that firms tend to issue positive news and withhold bad news (e.g. Kothari et al., 2009). Consistent with the same literature that firms withhold bad news until they cannot, bad news tends to be really bad – as Figure 2 shows that $\text{Neg}_\text{net}$, for example, of Dalian, is highly left-skewed.

Since the dictionaries overlap little with each other, the correlations of the sentiment measures across dictionaries are not high. As shown in Panel B of Table II, the correlation

<table>
<thead>
<tr>
<th>Panel A: summary statistics of the sentiment measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>$\text{Neg}_\text{net}$ (Dalian)</td>
</tr>
<tr>
<td>$\text{Neg}_\text{net}$ (HowNet)</td>
</tr>
<tr>
<td>$\text{Neg}_\text{net}$ (Common)</td>
</tr>
<tr>
<td>$\text{Neg}$ (Dalian)</td>
</tr>
<tr>
<td>$\text{Neg}$ (HowNet)</td>
</tr>
<tr>
<td>$\text{Neg}$ (Common)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: correlation matrix of the sentiment measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>(1) $\text{Neg}_\text{net}$ (Dalian)</td>
</tr>
<tr>
<td>(2) $\text{Neg}_\text{net}$ (HowNet)</td>
</tr>
<tr>
<td>(3) $\text{Neg}_\text{net}$ (Common)</td>
</tr>
<tr>
<td>(4) $\text{Neg}$ (Dalian)</td>
</tr>
<tr>
<td>(5) $\text{Neg}$ (HowNet)</td>
</tr>
<tr>
<td>(6) $\text{Neg}$ (Common)</td>
</tr>
</tbody>
</table>

Table II. Comparing sentiment measures across dictionaries

Notes: $\text{Neg}_\text{net}$ is the fraction of total negative word occurrences net of total positive word occurrences, and $\text{Neg}$ is the fraction of total negative word occurrences in each news article. “Common” refers to words that appear in all of the three dictionaries of Dalian, HowNet and NTU.
between \textit{Neg\_net (Dalian)} and \textit{Neg\_net (HowNet)} is 0.42, and that between \textit{Neg (Dalian)} and \textit{Neg (HowNet)} is 0.35. These correlations indicate that while sentiment measures across dictionaries are sufficiently different, these metrics, however, also tend to measure the sentiment of news articles consistently. We hence take comfort in using these metrics for our return regression exercises.

4.2 Is sentiment related to returns?

While there is a plethora of empirical associations documented from these sentiment measures in the literature, we link the sentiment measures to returns. Returns are supposedly the most informative measure – in fact, in designing their dictionaries, Yao et al. (2019) follow Engelberg et al. (2012) and use the cumulative abnormal returns following the releases of annual reports as the predictive information to “reverse-engineer” the Chinese sentiment word dictionary. The implicit assumption is that sentiment will drive returns, as shown in the literature, in particular, for the US market.

We link the sentiment to returns, retrieved from the Wind Financial Terminal, of the following horizons: previous day (day $[\neg 1]$), current day (day 0), the next day (day 1) and the next three days (days $[1, 3] \neg [4]$). We examine two returns: the stock’s return minus the market return (proxied by Wind All-A Stock Index return) – the market-adjusted return, or the stock return adjusted for returns on a comparable industry and size portfolio[5].

Table III presents the results of returns regressed on each of the sentiment measures. In Panel A, we do not impose on control variables; in Panel B, we include the control variables: the firm’s most recent $\beta$ (measured using the past three years’ monthly returns), the logarithm of market capitalization and the book-to-market equity ratio (e.g. Fama and French, 1996).

Since the evidence in Panels A and B is highly similar, we can focus our discussion on Panel A. The literature predicts a negative sign on \textit{Neg\_net} and \textit{Neg}; in particular, \textit{Neg\_net} and \textit{Neg} are able to negatively predict future returns (e.g. Tetlock et al., 2008; Huang et al., 2019).
### Table III
The association between sentiment and returns

<table>
<thead>
<tr>
<th></th>
<th>Market-adjusted return on day(s)</th>
<th>Industry- and size-adjusted return on day(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[−1]</td>
<td>0</td>
</tr>
<tr>
<td>Neg _net (Dalian)</td>
<td>0.005 [0.27]</td>
<td>0.002 [0.17]</td>
</tr>
<tr>
<td>Neg _net (HowNet)</td>
<td>−0.011 [−0.68]</td>
<td>−0.024 [−1.60]</td>
</tr>
<tr>
<td>Neg _net (Common)</td>
<td>−0.110** [−2.61]</td>
<td>−0.016 [−0.44]</td>
</tr>
<tr>
<td>Neg (Dalian)</td>
<td>−0.033 [−0.65]</td>
<td>−0.003 [−0.07]</td>
</tr>
<tr>
<td>Neg (HowNet)</td>
<td>−0.080* [−1.76]</td>
<td>0.074 [1.85]</td>
</tr>
<tr>
<td>Neg (Common)</td>
<td>−0.344*** [−2.25]</td>
<td>0.066 [0.49]</td>
</tr>
</tbody>
</table>

**Panel A: regressions without a control variable**

<table>
<thead>
<tr>
<th></th>
<th>Market-adjusted return on day(s)</th>
<th>Industry- and size-adjusted return on day(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[−1]</td>
<td>0</td>
</tr>
<tr>
<td>Neg _net (Dalian)</td>
<td>−0.015 [−0.96]</td>
<td>−0.01 [−0.69]</td>
</tr>
<tr>
<td>Neg _net (HowNet)</td>
<td>−0.01 [−0.63]</td>
<td>−0.034** [−2.23]</td>
</tr>
<tr>
<td>Neg _net (Common)</td>
<td>−0.104** [−2.59]</td>
<td>−0.012 [−0.34]</td>
</tr>
<tr>
<td>Neg (Dalian)</td>
<td>−0.02 [−0.42]</td>
<td>0.002 [0.05]</td>
</tr>
<tr>
<td>Neg (HowNet)</td>
<td>−0.073* [−1.68]</td>
<td>0.089** [2.24]</td>
</tr>
<tr>
<td>Neg (Common)</td>
<td>−0.284* [−1.95]</td>
<td>0.069 [0.52]</td>
</tr>
</tbody>
</table>

**Notes:** Each cell represents a regression and records the regression coefficient of stock return on Neg _net or Neg. Stock returns examined include market- or size-and-industry-adjusted returns on days [−1],[0],[1] and [1, 3], respectively. In Panel A, there is no control variables; and in Panel B, the control variables are β, log firm size and book to market. All variables other than Neg _net and Neg are omitted for brevity. t-Statistics are in square brackets. *,**,***Significant at 1, 5 and 10 percent levels, respectively.
Table III examines two future returns: days [1] and [1, 3]. For these two return horizons, we observe only a limited number of cases of $\text{Neg}_{\text{net}}$ and $\text{Neg}$ significance, but with mixed signs. For example, $\text{Neg}_{\text{net}}$ (Dalian) is positively related to returns of days [1] and [1, 3], yet $\text{Neg}$ (HowNet) is negatively related to returns of days [1, 3]. The evidence in Table III thus suggests that news sentiment in our sample is not likely to “correctly” predicts future stock returns.

News sentiment exhibits mixed signs on Day [0] returns, similar to the results for future returns. For example, $\text{Neg}_{\text{net}}$ (HowNet) and $\text{Neg}_{\text{net}}$ (Common) are negatively related to Day [0] industry- and size-adjusted return, but $\text{Neg}$ (HowNet) bears a positive sign. Taken together, Table III shows that there is limited evidence on the association between sentiment and either contemporaneous or future returns; and if there is any, the sign tends to flip, and the evidence does not warrant a strong conclusion.

Perhaps the most consistent evidence is from day [−1], where Table III shows that both $\text{Neg}_{\text{net}}$ and $\text{Neg}$ are negatively related to day [−1] returns. Take market-adjusted return in Panel A as an example: in the six cases presented there, three are significantly negative, and the rest three are insignificant. That is, not only does the mixed sign disappear, we also observe more cases of significance, consistent with the negative sign hypothesis.

At this point, it may be tempting to conclude the following based on the evidence in Table III: the day [−1] results suggest that media tends to follow significant stock returns and writes accordingly; and the contemporaneous and future return results suggest that media may even be facilitatory in pump-and-dump schemes by traders. That is, stock price jumps take place first, inducing media to write about the jumps, and traders in turn trade following media by engaging in profit taking. Apart from the limited sample that we use, a number of caveats to this interpretation, however, should be noted: the results only apply to the large stock set of SSE 50 stocks for a limited period of time, the news sequence in a sequentially covered event may distort the return prediction in a pooling sample, as demonstrated in Huang et al. (2019)[6] and press-initiated news may lag firm-initiated news, creating false return prediction by news.

Despite these caveats, the results in Table III point to a challenge in relying on contemporaneous and future returns as the “information” variable in reverse engineering a Chinese financial sentiment word dictionary. If contemporaneous and future returns do not convey information that are consistent with the sentiment in the texts, inferring a list of sentiment words from future stock returns, as done in Yao et al. (2019), would be biased. In our view, perhaps an alternative approach in such exercises is to normalize text sentiment in ways that take into account of news expectation (including media writing sequence), and link normalized sentiment to abnormal returns.

5. Discussion of directions for future research

We demonstrated in earlier sections that textual analysis in China’s financial markets is at a starting stage. We also showcased an application of sentiment analysis by studying news sentiment’s return associations of 2,500 news articles of SSE 50 firms. Given the importance of textual information in financial markets, we believe systematic textual research can create an impact on our understanding of the market. In our view, future research on Chinese financial texts can focus on the following areas.

The first area would be file parsing, tokenization and lexicons. Cross-sectional firm studies in finance require that text files to be comparable across firms and time. Oftentimes we convert files to plain texts for such purposes. One distinct feature of Chinese document files, in particular regulatory filings, are in the Portable Document Format (PDF) format. Parsing PDF files to more primitive format such as plain text or rich text format poses a challenge, as elements such as page layout, text boxes, tables and graphics often need to be taken care of. Fortunately, regulatory filings often follow specific format guidance and such guidance can be populated into limited formats. Nonetheless, we view parsing PDF files as a
first challenge in textual analysis in the Chinese market. A good textual analysis should design mechanisms to take into account of most of the PDF file variations for China's financial texts.

The next challenge in the first area in our view would be tokenization of financial texts. The Chinese language combines single characters into words with one to many syllables. While there exists some commonly used packages for Chinese word tokenization (such as the Jieba package that we used in our toy exercises), what we find in these packages is that they typically lack sufficient ability to distinguish financial entities (or entities in general) and phrases. Oftentimes financial entities and phrases contain many syllables and tend to be misclassified into multiple tokens, biasing the word count and general results that need to rely on word counts.

Another challenge is financial lexicon. Our toy application demonstrates the importance of financial sentiment word lexicons. There are a number of financial sentiment word lexicons, most notably the Chinese translation of Loughran and McDonald (2011) and Yao et al. (2019). These lexicons are, however, not without limitations, as discussed. As a sentiment lexicon is usually the starting point for common research applications, a well-defined lexicon is in urgent needs.

The second area that we believe that researcher can focus on is the sources of texts. Textual information in financial markets comes from either traditional media or social media. The rise of Chinese social media such as Weibo, Wechat and stock discussion forums, which all have a huge number of daily active users, naturally brings research attention to social media. We sift through a significant amount of stock discussion posts and find that information in these posts tend to contain little information. In other words, it is perhaps fair to argue that the "signal to noise" ratio in social media is low. Since Chinese financial textual analysis is at the starting phase, we advocate using more traditional textual sources such as regulatory filings, earnings announcements, corporate news, firm conference calls and analyst reports. These financial texts are from what we call "credible" sources, in that they contain high "signal to noise" ratio, as noise in these texts tend to be small. In contrast, we call social media "less credible" sources. As financial research is often tied to monetary decisions, we believe that researchers should focus on using credible sources to enhance the usefulness of their research in monetary decisions. If we are to use social media, we advocate that we should conduct a neck-to-neck comparison of research results with those from credible sources.

The third area for research is the measures of texts by applying new methodologies. At the current stage, Chinese financial textual analysis is focused on counting negative and positive words for sentiment gauging. Further to word counting, we can develop the degree of emotion (valence) for each word through, for example, machine learning methods. An extension can also be made to judge the sentiment of the corpus based on its thematic structure – for example, Azimi and Agrawal (2019) use a deep learning methodology to infer sentiment of sentences instead of words based on the US 10-K filings. Similar methodologies are used in Li et al. (2019). Researchers can further develop suitable topic modeling methodologies for Chinese texts. Other than bag of words and machine learning classification methods, we should develop tools to extract huge information from financial texts. For example, Calomiris and Mamaysky (2019) recently measure the unusualness of word flow using “entropy” (measuring “disordering” of texts) and apply it to global risk-return trade-off.

Last but not least, we raise a number of research applications that we view of interest. We can classify finance research to be in the broad areas of asset pricing and corporate finance. Coupled with the Chinese setting, some interesting questions in these two areas may emerge. Directly related to the experiment we conducted, for example, what are the return implications – past and future returns, return volatility and other return patterns – of
different classes of financial texts, such as annual reports, news and firm announcements? Moreover, it is well noted that Chinese markets closely tied to government public policies (e.g. Brunnermeier et al., 2017). Given so, what are the implications of government policy announcements? Are we able to classify the above texts into different broad-class of topics and study the cross-sectional differences and impacts of the topics? Do market participants exhibit divergence of opinions and exert the divergence on their portfolio holdings? Do market reactions to firm announcements or firm news induce firm management to change their behaviors? Slightly moving away from stock market performance, effectively controlling financial risk, e.g., leverage, systemic risk and security comovement, is a major concern in the China’s financial system (see e.g. Liu et al., 2018).

The world today assimilates unstructured data much better and faster than any other time in our history. Much of such assimilation now, we believe, hinges on critical domain knowledge. Many novel research ideas have come out of quantitative structured data in finance in the past 50 years. We believe that the next decade will see significant insights gained from our understanding of the unstructured financial text data. Researchers armed with ideas, finance domain knowledge and NLP tools are in a sweet position to deliver exciting findings on Chinese financial texts.

Notes
2. For the word lists, see https://sraf.nd.edu/textual-analysis/resources/
3. We went through the first 20 pages of search results on the phrase “Hong Kong finance textual analysis” on both Google and Google Scholar, and also went through Hong Kong Monetary Authority’s Research Memorandums series and working papers series in the past seven years. The only paper that bears relevance to this project is Li et al. (2014), who use news stories of 22 Hong Kong stocks based on Loughran and McDonald’s (2011) lexicon.
4. Returns and news are aligned by the time stamp of news. Day 0 is the (trading) day of the news if news takes place pre-market or intra-day before market closes, and is the next trading day if news takes place post-market.
5. The industry and size portfolio is constructed using the first-letter-industry classification by Chinese Securities Regulatory Commission, coupled with a tercile breakdown of market capitalization for each industry.
6. To see this, imagine that an event is covered consecutively by media in days [0] and [1], and returns on both days consistently reflect negatively the news sentiment. In a pooling regression, it is likely that day [0] return is related to day [1] news sentiment, resulting, inadvertently, in a negative coefficient of day [1] news sentiment on day [0] return.

References


**Corresponding author**

Alan Huang can be contacted at: aghuang@uwaterloo.ca

For instructions on how to order reprints of this article, please visit our website: [www.emeraldgrouppublishing.com/licensing/reprints.htm](http://www.emeraldgrouppublishing.com/licensing/reprints.htm)

Or contact us for further details: permissions@emeraldinsight.com