Financial Innovation

RESEARCH Open Access

Tone of language, financial disclosure, and earnings management: a textual analysis of form 20-F

Shuangyan Li[1], Guangrui Wang[2] and Yongli Luo[3*]

*Correspondence:
yluo@hbu.edu
[3] Archie W. Dunham College of Business at Houston Baptist University, Houston, USA
Full list of author information is available at the end of the article

Abstract

This study investigates the relationship between the tone of financial disclosures and managers' earnings management behavior using Form 20-F filings of Chinese firms listed in the U.S. during 2002–2014. The results show that the proportion of positive, uncertain, or modal words used in financial disclosures is positively related to corporate earnings management, implying that managers attempt to conceal earnings management behavior by manipulating the tone of their financial reports. In addition, robustness tests are conducted using an alternative proxy for earnings management that considers the effects of the financial crisis and separately examines the information and non-information technology industries. The results suggest that the tone used in financial disclosures has informative value, and textual analysis can be an effective tool for identifying earnings management.

Keywords: Tone, Earnings management, Textual analysis, Financial statement

JEL Classification: G15, M41, G15

Introduction

The U.S. Securities and Exchange Commission (SEC) requires all non-U.S. companies engaging in securities trading in the U.S. to file an annual Form 20-F. Recently, the SEC has finalized rules to implement the Holding Foreign Companies Accountable Act (HFCAA), that allow the delisting of foreign firms if their auditors do not comply with requests for information from the U.S. regulators. The goal of this requirement is to standardize the reporting of foreign-based companies and inspect the audits of Chinese firms that list and trade in the U.S. market. It has been documented that the quality and accuracy of financial information reported by foreign firms cross-listed in the U.S. is low (Ang et al. 2016). Furthermore, Beckmann et al. (2019) claim that foreign firms have incentives to manage their earnings around cross-listing events. For example, the Chinese firm, Luckin Coffee, had an internal fraud probe in 2020, increasing calls for action. In addition, due to information asymmetry and window-dressing in cash holdings (Khokhar 2011), there is a significant positive association between qualitative impression management and abnormal accruals in earnings management (Boudt and

Thewissen 2019), and executives likely attempt to hide their earnings management activities in these disclosures (Jaspersen et al. 2021). Therefore, it is imperative for investors, analysts, and legislators to make special efforts to identify earnings management behaviors in foreign firms'20-F financial statements.

The Form 20-F consists of several sections, including key information, information on the firm, operating and financial prospects, directors, senior management, and employees. The SEC requires all "foreign private issuers" with listed equity shares on the U.S. exchanges to provide investors with detailed information about a firm's annual business activities. Therefore, it is assumed that texts conveying uncertainty make it more challenging for investors to identify earnings management behaviors. Additionally, the tone of such documents may impact investors' assessments of the firm (Loughran and McDonald 2016). Although the 20-Fs indicate that firms follow the U.S. rules and offer opportunities for investors on the U.S. exchanges, the tone of the financial statements may be different from the tone used by the U.S. companies. This study adopts a novel Chinese dataset because Asian cultures have more implicit communication styles than nonverbal cues (Wu et al. 2021). Moreover, Chinese individuals are more likely to hide their true feelings than Americans because of the cultural characteristics of collectivism (Sims et al. 2015). Particularly, Chinese stocks play an important role in the global economy in international portfolio diversification (Luo et al. 2012). However, understanding the true meaning of the language used in Chinese companies' financial reports is challenging for international investors because the tone of language represents the underlying corporate culture. Specifically, firm performance dimensions are positively affected by organizational culture and innovation (Uzkurt et al. 2013). In general, this study attempts to identify earnings management behaviors in the U.S.-listed Chinese companies by analyzing the qualitative and quantitative information contained in the 20-F disclosures.

The existing literature identifies and focuses on factors affecting earnings management from both macro and micro perspectives. Studies include state-level unemployment benefits (Dou et al. 2016), trust culture (Pevzner et al. 2015), controlling ownership (Bao and Lewellyn 2017), dual-class status (Li and Zaitas 2017), corporate governance (El Diri et al. 2020), and stakeholder orientation (Ni 2020). However, there is a lack of research on how investors can use effective tools to analyze financial disclosures, such as Form 20-Fs, to identify earnings management behaviors. This study contributes to the literature by applying existing dictionary-based textual analysis methods to analyze the U.S.-listed Chinese companies.

Another stream of research examines the relationship between the textual characteristics of financial disclosures and the financial market's evaluations of public firms (Loughran and McDonald 2013; Lang and Stice-Lawrence 2015). These studies focus on the various outcomes of textual analysis, but there are few studies on how textual analysis can help identify earnings management behavior. This study applies textual analysis to the U.S.-listed Chinese firms because an individual country sample can support an in-depth analysis of the business culture characteristics in that country. Textual analysis can effectively identify the tone and sentiment expressed in mandated company disclosures, analyst reports, and financial press releases. Loughran and McDonald (2016) find that the language used in financial statements, especially the tone and words indicating

sentiment, correlate with earnings management and may signal future fraudulent activities. Several studies use textual analysis to quantify the tone and sentiment of the language used in corporate annual reports. They examine the association between the tone of the language and performance outcomes and market interpretations of such information (Loughran and McDonald 2011, 2013; Jegadeesh and Wu 2013; Lang and Stice-Lawrence 2015). Moreover, the tone used in the Management Discussion and Analysis (MD&A) section in financial disclosures is highly associated with market responses to publicly listed firms (Wu et al. 2021).

This study investigates the relationship between the tone of 20-F and earnings management behaviors of Chinese companies listed in the U.S. exchanges. Corporate managers typically have an information advantage over outside investors, causing information asymmetry and window-dressing problems in cash holdings (Khokhar 2011). Specifically, Jegadeesh and Wu (2013) indicate a significant relationship between document tone and market reactions for both negative and positive words. Recently, Bian et al. (2021) confirm that management tone is significantly and positively related to a firm's post-market operating performance. Therefore, we examine whether managers have an incentive to select their textual tone in financial reports when communicating company information to outsiders.

It is crucial to adopt an appropriate method to measure the tone of 20-Fs. Previous literature shows that the Loughran and McDonald word list is widely used in textual analyses (Jegadeesh and Wu 2013; Loughran and McDonald 2013). This study adopts the six sentiment word lists (positive, negative, uncertain, litigious, strong modal, and weak modal) following Loughran and McDonald (2011). It adds two more words (moderate modal and irregular verbs) from McDonald's updated 2014 Master Dictionary, providing a broader set of sentiment categories and more examples of words for each category.[1] Specifically, this study adds a moderate modal list because the modal subcategories are too broad, presenting a complete picture of the relationship between modal words and earnings management. It adds a list of irregular verbs because it is believed that these words can reflect major events occurring in the company and thus convey additional information on the most recent mergers and acquisition activities.

This study extends Kim et al. (2017) and Loughran and McDonald (2011). For example, Kim et al. (2017) examine the relationship between the grammatical structure of texts and financial reporting characteristics, separating texts into two types based on how they encode time: strong future-time reference texts and weak future-time reference texts. In contrast, this study conducts a textual analysis of the tone of Form 20-Fs filed by the U.S.-listed Chinese firms and examines whether the tone correlates with earnings management, distinct from other accounting characteristics. Specifically, this study examines whether a higher proportion of positive, uncertain, or modal words like "good, may, could, depend, and approximately" in the 20-Fs signals more earnings management behaviors.

This study applies ordinary least squares (OLS) regression models to a sample of 153 U.S.-listed Chinese firms during 2002–2014. The results show that Chinese firms with

[1] The word lists can be found at the following website https://sraf.nd.edu.

a higher proportion of positive, uncertain, or modal words in their 20-Fs exhibit higher earnings management, suggesting that managers use positive, uncertain, and modal terms to hide their earnings management behaviors. A series of robustness tests are also conducted. First, an alternative proxy for earnings management is used as the dependent variable, and the model is re-estimated; the main results hold. Second, an examination of the impact of the 2008 financial crisis reveals that the relationship between tone and earnings management is more significant in the post-crisis period than in the pre-crisis period. Third, a robustness test of cross-industry variations finds that the relationship between tone and earnings management is more significant in the information technology industry than in other sectors.

This study makes two significant contributions to the literature. First, it adds the Chinese context to the current textual analysis literature (Loughran and McDonald 2011, 2013; Price et al. 2012; Jegadeesh and Wu 2013; Kim et al. 2017; Shan 2019; Wu et al. 2021) and the predictability of qualitative disclosures on earnings management (Hu 2021; Jaspersen et al. 2021). This is the first study using the U.S.-listed Chinese firms to investigate the correlation between textual tone and earnings management behaviors. The findings provide novel empirical evidence that managers tend to use positive, uncertain, or modal words in financial statements to conceal their earnings management behaviors, supported by Li's (2008) obfuscation theory and Boudt and Thewissen's (2019) impression management. Second, this study contributes to the literature on the methodology used to identify potential earnings management. For example, Li et al. (2021a) claim that in many financial applications, such as fraud detection, reject inference, and credit evaluation, it is critical to understand the sub-patterns of the data to infer users' behaviors, identifying potential risks. Specifically, this study illustrates that the textual tone in financial statements can be used as an effective tool to identify earnings management behaviors because empirical studies show the correlation between tones and earnings management, providing informative value for identifying earnings management behavior. However, previous studies highlight the correlation between investor relations, ownership, corporate governance, firms' financial characteristics, shareholder activism, environmental regulation, and corporate culture with earnings management (Frankel et al. 2010; Bao and Lewellyn 2017; Dou et al. 2016; El Diri et al. 2020; García Lara et al. 2020; Ni 2020; Chen et al. 2021; Li et al. 2021b; Ng et al. 2021).

The remainder of this paper is organized as follows. In "Literature review and hypothesis development" section presents a comprehensive literature review and develops our main hypotheses. In "Data and methodology" section introduces the data and methodology. In "Results" section reports the empirical analysis results, and in "Conclusions" section concludes the paper.

Literature review and hypothesis development

Opinion dynamics in finance and business have been studied extensively, and a series of opinion dynamics models with different opinion evolution rules have been proposed in various fields. Opinion dynamics models typically include three basic elements: opinion expression formats, opinion evolution rules, and opinion dynamics environments (Zha et al. 2020). Textual analysis extracts information from textual resources from a new perspective (Loughran and McDonald 2016). Therefore, the tone in the text is used

to represent investor sentiment and analyze issues in accounting and finance behaviors. Specifically, it helps us understand behavioral patterns across individuals, institutions, and markets (Kearney and Liu 2014; Yang and Luo 2014). In addition, these techniques may help researchers analyze hidden clues or seek additional information to that observed through financial information, increasing the quantity and quality of information (Gandía and Huguet 2021) and creating different methodologies and artifacts to advance knowledge in accounting, auditing, and finance (Fisher et al. 2016). Numerous studies examine the association between textual tone or textual characteristics and firm performance and the market reaction to the tone of financial disclosures. For example, Loughran and McDonald (2014) find that file size (in megabytes) is an accurate indicator of a text's readability and is related to post-filing abnormal returns, volatility, and analyst dispersion. Bodnaruk et al. (2015) use the percentage of constraining words in financial statements to gauge firms that would become financially constrained and find that this measure is associated with liquidity events. Loughran and McDonald (2013) use sentiment word lists to gauge the tone used in the initial registration forms of new securities and find that initial public offerings (IPOs) with high levels of uncertain words have higher first-day returns and larger aftermarket volatility. In addition, Jegadeesh and Wu (2013) propose a return-based term-weighting scheme for quantifying document tone and find that the tone of financial statements is significantly related to the market returns of firms around their financial statement filing dates. Mai et al. (2019) assess the predictive power of textual data for forecasting bankruptcy and find that textual data can complement traditional accounting- and market-based variables in predicting bankruptcy. Wei et al. (2019) adopt the text mining approach to identify corporate energy risk factors in the risk disclosures reported in financial statements and provide a preliminary method for corporate energy risk estimation. Jiang et al. (2019) construct a manager sentiment index based on the aggregated textual tone of corporate financial disclosures and find that manager sentiment is a strong negative predictor of stock market returns.

It is well documented that people's opinions drive human decisions and behaviors, and some information in their environment is always reflected in their language (Hofstede and Hofstede 2010). People express their opinions and sentiments through oral or written communication (Abbasi et al. 2008). For example, people may use vague statements to express their views or a strong tone to deny negative behaviors. In business environments, managers have an information advantage over outside investors. Communication about the firm's business and financial performance may partially reflect managers' subjective opinions about their stock returns, financial matters, fraudulent activities, or future earnings expectations (Henry 2008; Blau et al. 2015; Demaline 2019; Jiang et al. 2019). For example, Jegadeesh and Wu (2013) find that positive and negative tones in Form 10-K are useful in predicting the date of filing returns. Bian et al. (2021) find that managers' positive tone in online roadshows can indicate a higher first-day return on IPOs in Chinese stock markets, suggesting that the sentiment or tone of managers has informative value for investors in certain events or activities. In addition, Jaspersen et al. (2021) argue that managers may be more likely to hide some earnings management behaviors in qualitative disclosures, making it difficult for investors to make correct decisions. This is consistent with impression management theory, which states that qualitative impression management positively correlates with the use of abnormal accruals

in earnings management (Boudt and Thewissen 2019). Therefore, it is predicted that a significant relationship exists between sentiment expressed in the language of these texts and earnings management. Specifically, this study hypothesizes that:

Hypothesis 1 The tone used in Form 20-Fs is significantly correlated with managers' earnings management behavior.

A natural question arises: How does tone relate to earnings management? A higher proportion of positive, uncertain, or modal words such as "good, may, could, depend, and approximately" reflect good or uncertain information about the firm, suggesting that managers convey accurate information to the public. In this situation, it can be inferred that managers do not conceal earnings management. However, managers may also use a positive or uncertain tone to hide their earnings management behavior. Studies show that the annual reports of public firms with low earnings can be difficult to read (Li 2008) and the Chinese firms with more influential largest shareholders are more prone to real earnings management (Dong et al. 2020), indicating that managers may be reluctant to report accurate information when the news is bad. Kang et al. (2018) find that managers use a flexible tone in the narrative sections of annual reports to express their specific intentions according to different firm situations. Specifically, managers may strategically use language when a firm violates regulations (Huang et al. 2014). Managers may also attempt to deceive investors out of self-interest or hide bad news that may damage the firm's value (Kang et al. 2018). Accordingly, textual analysis can reveal firms' fraudulent activities (Chen 2014). Jaeschke et al. (2018) find that the tone of financial reports correlates with the likelihood of future prosecution under the *Foreign Corrupt Practices Act* (FCPA) and that after the prosecution, the FCPA violators tend to change the tone in their financial reports. Therefore, we construct an alternative hypothesis as follows:

Hypothesis 1A A positive, uncertain, or modal tone in 20-Fs positively correlates with managers' earnings management behavior.

Data and methodology

Data

The sample is obtained from the S&P Capital IQ with supplements from the Wind Database and the China Stock Market & Accounting Research Database (CSMAR). The initial sample includes 153 Chinese non-financial firms[2] listed on the National Association of Securities Dealers Automated Quotations (NASDAQ), New York Stock Exchange (NYSE), and American Stock Exchange (AMEX) from 2002 to 2014. Next, companies that released 10-K filings[3] without any other SEC filings are removed. Finally, filings with incomplete financial information are excluded. The final sample consists of 75 companies and 449 firm-year observations.

[2] The financial firms are excluded because the high leverage that is normal for these firms does not have the same meaning as for non-financial firms, where high leverage usually indicates distress (Fama and French 1992).

[3] Form 10-K is a comprehensive report of financial performance filed annually by the U.S. public firms.

Word lists

Following the 2014 Master Dictionary of McDonald and the Loughran and McDonald (2011) word lists, this study classifies the words into eight categories: positive (*positive*), negative (*negative*), uncertain (*uncertainty*), litigious (*litigious*), strong modal (*strongml*), moderate modal (*moderml*), weak modal (*weakml*), and irregular verbs (*irrverb*). This study uses the following examples for each category to clarify the definitions of these sentiment words: examples of positive words are "beneficial, successful, good, achieved, and empower"; examples of negative words are "loss, failure, abandon, and decline"; examples of uncertainty words are "almost, and nearly"; examples of litigious words are "abovementioned, abrogated, and certiorari"; examples of strong modal words are "must, never, definitely, and will"; examples of moderate modal words are "can, generally, and usually"; examples of weak modal words are "could, should, and ought to"; and the examples of irregular verbs are "beat, cut, and forgot." Keywords extraction plays an important role in the financial sector. "With this combination of keyword extraction and big data, we can extract what we need in a very short time with a fully automated process eliminating most manual work and increasing speed of data collection (Pejic Bach et al. 2019)."

The textual analysis method applies to 10-K filings by industry, as described in Appendix 2 (Bodnaruk et al. 2015). Initially, this study removes all the HTML and ASCII-encoded segments from each filing; then, the text identified by the HTML as tables is eliminated if its numeric character content is greater than 15%. After removing ambiguous proper nouns, the text is parsed into a vector of words tabulated using the eight-word list categories. It calculates the percentage of words in each category using the approach proposed by Loughran and McDonald (2013).

Variables

This study estimates discretionary revenues as a proxy for earnings management, following McNichols and Stubben (2008) and Stubben (2010). Each approach to identifying earnings manipulation has advantages and disadvantages (McNichols and Stubben 2008). Discretionary accrual measures are commonly used to identify firms that engage in earnings management. This measure does not involve selection bias but may have a greater measurement error. Several studies have shown that discretionary accrual models provide biased, low-power estimates of discretion. We use the measure of discretionary revenues presented by Stubben (2010) as a proxy for earnings manipulation to increase the power of our tests. Specifically, we run the following regression in Eq. (1):

$$\Delta AR_{i,t} = \alpha_0 + \alpha_1 \Delta Rev_{i,t} + \varepsilon_{i,t} \tag{1}$$

where the subscripts refer to firm i in year t. $\Delta AR_{i,t}$ represents the annual change in accounts receivable and $\Delta Rev_{i,t}$ is the annual change in revenues; both are scaled by lagged total assets. Discretionary revenues are the residuals estimated using Eq. (1). The absolute values of the discretionary revenue are then called earnings management (EM).[4]

[4] This study uses absolute values because it emphasizes the occurrence of earnings management behavior rather than the direction of earnings management.

This study includes various control variables that previous studies have found to affect *E.M*. First, it uses a dummy variable to control for the listing place (*LP*), which equals 1 if the firm is listed on the NASDAQ and 0 otherwise. The variable for firm nature (*FN*) is set to 1 if the controlling shareholder is the state and 0 otherwise, because state control may increase earnings management (Bao and Lewellyn, 2017). It creates a dummy variable for the presence of an auditor going-concern (*AOT*), which equals 1 if an auditor has issued a going-concern opinion and 0 otherwise. It is assumed that when auditors issue going-concern opinions firms have less incentive to engage in earnings management. It further controls for auditing firm characteristics with the variable (*BIG4*), which equals 1 if a firm is audited by one of the Big 4 auditing companies and 0 otherwise (Hu 2021) because these auditors may exercise stronger supervision, resulting in fewer earnings management. We also control for the following firm characteristics: Leverage (*LEV*), defined as the ratio of total debt to total assets, indicates that firms with high leverage manage earnings to avoid debt covenant violations. Firm size (*LNSIZE*) is the natural logarithm of a firm's total assets and is used to proxy for size effects. Most studies control for this variable, but their results are inconsistent (Hu 2021; Ng et al. 2021). In addition, we control for cash and cash equivalents (*CH*), defined as the ratio of cash and cash equivalents to total assets, and revenue (*REV*), measured as the change in sales the lagged sales of the company. Cash flow (*CF*) is defined as the cash flow ratio from operations to total assets. It is assumed that firms with high cash flow engage in fewer earnings management because good cash flow represents strong operations. This also includes return on assets (*ROA*), measured by net income over total assets. However, the sign of its impact on earnings management is uncertain, because firms with good and bad financial performances may engage in earnings management. Finally, we include firm, year, and industry fixed effects to control for heterogeneity. Appendix 1 summarizes the variable definitions.

Methodology

The following OLS regression is employed to test the main hypotheses and controls for the firm, year, and industry fixed effects in Eq. (2):

$$EM_{it} = \beta * WordLt_{it} + \gamma * X_{it} + Firm_i + Year_{it} + Industry_j + \varepsilon_{it} \quad (2)$$

where *i*, *t*, and *j* refer to firm, year, and industry, respectively. *EMit* represents the earnings management of firm *i* in year *t*, β is a set of coefficients; $WordLt_{it}$ represents the set of eight sentiment word lists; and X_{it} represents the control variables (LP_{it}, FN_{it}, AOT_{it}, LEV_{it}, $LNSIZE_{it}$, CH_{it}, REV_{it}, and CH_{it}). $Year_{it}$ represents year-fixed effects and $Industry_j$ represents industry-fixed effects.

Furthermore, this study implements a multicollinearity test for the main regression model. In addition, it uses an alternative proxy for earnings management and re-estimates Eq. (2). Finally, to test the robustness of the empirical results, we divide the full sample into two sub-samples, each with different periods and industries. This study also uses the fixed effects model in Eq. (2) to test the impact of the 2008 financial crisis on the relationship between independent variables and earnings management and the cross-industry variations between information and non-information technology industries.

Table 1 Summary of descriptive statistics

Variable	Mean	Mean (Year < 2008)	Mean (Year ≥ 2008)	Median	Minimum	Maximum	SD
EM	0.030	0.025	0.031	0.017	0.000	0.341	0.040
Positive	0.003	0.004	0.003	0.003	0.000	0.008	0.001
Negative	0.007	0.009	0.007	0.007	0.001	0.017	0.003
Uncertainty	0.006	0.007	0.006	0.005	0.002	0.012	0.002
Litigious	0.007	0.009	0.006	0.006	0.001	0.030	0.004
Strongml	0.002	0.002	0.002	0.002	0.000	0.004	0.001
Moderml	0.001	0.002	0.001	0.001	0.000	0.003	0.001
Weakml	0.003	0.004	0.003	0.003	0.001	0.007	0.002
Irrverb	0.003	0.005	0.003	0.004	0.001	0.009	0.001
LP	0.490	0.005	0.493	0	0	1	0.500
FN	0.350	0.605	0.289	0	0	1	0.477
AOT	0.976	0.988	0.973	1	0	1	0.155
BIG4	0.846	0.919	0.829	1	0	1	0.361
LEV	0.387	0.322	0.403	0.355	0.021	0.952	0.209
LNSIZE	3.034	3.125	3.012	2.779	1.025	5.589	0.927
CH	0.247	0.300	0.234	0.173	0.003	0.898	0.223
REV	0.285	0.431	0.251	0.204	−0.860	4.064	0.480
CF	0.086	0.137	0.074	0.081	−0.441	0.442	0.111
ROA	0.032	0.077	0.021	0.036	−0.555	0.368	0.108

This table reports the summary of descriptive statistics of the sample including 75 U.S.-listed Chinese companies during 2002–2014.The total number of observations is 449. The sample is obtained from the S&P Capital IQ and supplemented from the Wind Database. LP is a dummy for listing place, FN is a dummy for state-owned enterprise, AOT is a dummy for auditor going concern, and BIG4is a dummy for Big 4 auditing firms. LEVis the ratio of total debt to total assets. LNSIZE is the logarithm of firm's total assets. CH is the ratio of cash to total assets. REV is the change in sales. CF is the ratio of cash from operating to total assets. ROA is return on assets. Positive, negative, uncertain, litigious, strong modal, moderate modal, weak modal, and irregular verbs are the eight categories of wordlist defined from the 2014 Master Dictionary of McDonald and Loughran and McDonald (2011) word list

Results

Summary statistics

The industry subsamples are based on the Global Industry Classification Standard (GICS) developed by the S&P Dow Jones Indices, a leading provider of global equity indices. After excluding the financial industry (see Appendix 2), there were nine industries in the sample. Over the entire sample period, the information technology industry accounts for 42.98% of the total observations, while the GICS utility industry accounts for only 1.78%. The weights of the other industries vary from 2.67% (communication services) to 16.7% (consumer discretionary). In general, sample firms are widely dispersed across various industries.

Table 1 reports the summary statistics for the selected variables. For example, for the dependent variable, *EM*, the minimum is 0, and the maximum is 0.341, with a mean of 0.031 and a median of 0.017. These results suggest variations in the amount of earnings management in the sample firms. Table 1 also reports the descriptive statistics of the eight sentiment measures, the percentages of words in Form 20-F selected from each word list. The mean values are greater than the median values for most sentiment variables, implying that the data distribution is skewed and outliers. Specifically, *negative*, *litigious*, and *uncertain* have the highest mean percentages (0.007, 0.007, and 0.006, respectively). In addition, it reveals that firm size (*LNSIZE)*

and change in sales (*REV*) have the highest standard deviations, 0.927 and 0.480, respectively, suggesting that the sample firms' revenues vary remarkably by firm size.

This study also examines the effect of the 2008 financial crisis on the selected variables. Specifically, it divides the sample into two sub-samples: a sub-sample of 86 observations during 2002–2007 and a sub-sample of 363 observations during 2008–2014. This study uses the beginning of 2008 as the cut-off point for the financial crisis. It compares the means of selected variables before and after 2008. The mean *EM* is slightly higher in the post-crisis sub-sample than in the pre-crisis period (0.031 vs. 0.025). We also find that the values for all sentiment variables are lower in the post-crisis period. For example, the mean of *litigious* drops from 0.009 to 0.006. Finally, the results show that the debt ratio (0.403 vs. 0.322) increases, while the cash ratio (0.234 vs. 0.300), cash flow ratio (0.074 vs. 0.137), and change in sales (0.251 vs. 0.431) decrease after the financial crisis. The results suggest that most firms have experienced financial constraints during the economic problems.

Table 2 reports the correlations between key variables. Most sentiment variables are positively associated with *EM*, except for *litigious*. For example, *Positive*, *Uncertainty*, *Strongml*, *Moderml*, and *Weakml* are positively correlated with *EM* (the coefficients are 0.196, 0.169, 0.092, 0.132, and 0.170, respectively), although the correlations between *EM* and *Negative* (0.029) and *Irrverb* (0.028) are weak. These results are consistent with those of Loughran and McDonald (2013), although there is an overlap between some categories in the word list. For example, there are strong correlations between *Weakml* and *Uncertainty* (0.954), *Uncertainty* and *Positive* (0.895), and *Moderml* and *Positive* (0.887), suggesting that multicollinearity will be a problem if it includes all the sentiment variables in one regression. Therefore, this study follows Loughran and McDonald (2013) and runs the regressions individually for each category of sentiment variables. In addition, there is a high correlation coefficient between *CF* and *ROA* (0.755), suggesting that firms with higher cash flows are more likely to have better performance. A possible explanation is that these firms have lower fixed asset costs and more capital turnover, resulting in higher profits.

Baseline regression results

To further examine the link between earnings management and sentiment, we first report the baseline regression results, including the firm, year, and industry fixed effects in Table 3. Column (1) shows that the correlation between firm nature and earnings management is positive (0.013) and significant at the 5% level. This demonstrates that the managers of state-owned enterprises engage in more earnings management, consistent with studies of the Chinese financial markets (Bao and Lewellyn 2017). The firm size and cash flow coefficients are negative and significant, suggesting that larger firms with higher cash flows have fewer earnings management activities. In other words, small firms and firms with inadequate cash flow are more likely to manipulate their earnings. The coefficient of firm leverage is positive and significant, suggesting that firms with high debt ratios engage in earnings management, consistent with the original assumption.

Columns (2)–(9) of Table 3 examine each of the eight sentiment categories as independent variables to explore whether the language used in financial disclosures can explain earnings management. The results show that most sentiment variables (*Positive*,

Table 2 Correlation matrix

	EM	Positive	Negative	Uncertainty	Litigious	Modal1	Modal2	Modal3	Irr_verb	LEV	LNSIZE	CH	REV	CF	ROA
EM	1														
Positive	0.196***	1													
Negative	0.029	0.763***	1												
Uncertainty	0.169***	0.895***	0.756***	1											
Litigious	-0.009	0.420***	0.767***	0.344***	1										
Strongml	0.092*	0.841***	0.750***	0.807***	0.457***	1									
Moderml	0.132**	0.887***	0.732***	0.883***	0.364***	0.846***	1								
Weakml	0.170***	0.869***	0.787***	0.954***	0.419***	0.745***	0.830***	1							
Irrverb	0.028	0.668***	0.757***	0.670***	0.695***	0.747***	0.689***	0.607***	1						
LEV	− 0.004	− 0.197***	− 0.235***	− 0.261***	− 0.174***	− 0.191***	− 0.272***	− 0.284***	− 0.115**	1					
LNSIZE	0.208***	− 0.258***	− 0.179***	− 0.219***	− 0.042	− 0.138**	− 0.263***	− 0.339***	0.106**	0.417***	1				
CH	0.008	0.199***	0.337***	0.220***	0.294***	0.181***	0.199***	0.324***	0.079*	− 0.435***	− 0.523***	1			
REV	0.162***	0.087**	0.108**	0.029	0.172***	0.052	0.048	0.077	0.113**	0.019	− 0.061	0.119**	1		
CF	− 0.135***	0.026	0.032	0.022	0.110**	0.089*	0.050	0.016	0.148**	− 0.086**	0.254***	0.124**	0.236***	1	
ROA	0.049	0.119**	0.009	0.106**	0.055	0.137***	0.121**	0.095	0.150	− 0.192	0.147**	0.111**	0.251***	0.755***	1

This table reports the correlation matrix of selected variables. The sample includes 75 U.S.-listed Chinese companies during 2002–2014.The total number of observations is 449. The sample is obtained from the S&P Capital IQ and supplemented from the Wind Database. LP is a dummy for listing place, FN is a dummy for state-owned enterprise, AOT is a dummy for auditor going concern, and BIG4is a dummy for Big 4 auditing firms. LEVis the ratio of total debt to total assets. LNSIZE is the logarithm of firm's total assets. CH is the ratio of cash to total assets. REV is the change in sales. CF is the ratio of cash from operating to total assets. ROA is return on assets. Positive, negative, uncertain, litigious, strong modal, moderate modal, weak modal, and irregular verbs are the eight categories of wordlist defined from the 2014 Master Dictionary of McDonald and Loughran and McDonald (2011) word list

Table 3 OLS Regression on earnings management

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent variable: EM									
Positive		6.482***							
		(3.84)							
Negative			1.113						
			(1.48)						
Uncer-tainty				3.793***					
				(4.00)					
Litigious					0.406				
					(0.87)				
Strongml						6.372**			
						(2.25)			
Moderml							8.741**		
							(2.18)		
Weakml								5.357***	
								(3.60)	
Irrverb									3.185
									(1.63)
LP	−0.007	−0.005	−0.006	−0.004	−0.006	−0.006	−0.005	−0.004	−0.005
	(−1.39)	(−1.03)	(−1.22)	(−0.90)	(−1.35)	(−1.28)	(−1.10)	(−0.80)	(−1.10)
FN	0.013**	0.014**	0.014**	0.013**	0.014**	0.011*	0.012*	0.015**	0.012*
	(2.02)	(2.13)	(2.10)	(2.07)	(2.05)	(1.66)	(1.82)	(2.36)	(1.83)
AOT	0.013	0.014	0.013	0.013	0.014	0.012	0.012	0.014	0.013
	(1.06)	(1.17)	(1.06)	(1.09)	(1.10)	(1.01)	(0.97)	(1.17)	(1.04)
BIG4	−0.003	−0.001	−0.003	−0.001	−0.003	−0.003	−0.003	−0.001	−0.003
	(−0.44)	(−0.10)	(−0.47)	(−0.11)	(−0.50)	(−0.41)	(−0.37)	(−0.20)	(−0.51)
LEV	0.017	0.016	0.018*	0.020*	0.018	0.016	0.018*	0.019*	0.018
	(1.57)	(1.46)	(1.66)	(1.86)	(1.61)	(1.49)	(1.67)	(1.75)	(1.60)
LNSIZE	−0.018***	−0.016***	−0.018***	−0.017***	−0.018***	−0.017***	−0.016***	−0.016***	−0.018***
	(−3.69)	(−3.23)	(−3.72)	(−3.46)	(−3.71)	(−3.43)	(−3.30)	(−3.40)	(−3.66)
CH	−0.009	−0.009	−0.012	−0.011	−0.011	−0.011	−0.009	−0.011	−0.011
	(−0.78)	(−0.78)	(−0.99)	(−0.93)	(−0.90)	(−0.94)	(−0.78)	(−0.91)	(−0.92)
REV	0.015***	0.016***	0.015***	0.017***	0.015***	0.016***	0.016***	0.016***	0.015***
	(3.61)	(3.70)	(3.56)	(3.93)	(3.45)	(3.76)	(3.73)	(3.83)	(3.49)
CF	−0.119***	−0.110***	−0.119***	−0.110***	−0.119***	−0.117***	−0.117***	−0.111***	−0.118***
	(−4.44)	(−4.16)	(−4.42)	(−4.14)	(−4.44)	(−4.36)	(−4.38)	(−4.19)	(−4.41)
ROA	0.122***	0.107***	0.125***	0.109***	0.123***	0.116***	0.116***	0.109***	0.122***
	(4.47)	(3.95)	(4.58)	(4.04)	(4.52)	(4.28)	(4.25)	(4.01)	(4.48)
Firm	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	449	449	449	449	449	449	449	449	449
Adj R^2	0.136	0.164	0.139	0.166	0.136	0.145	0.144	0.161	0.140
F-statistics	5.53	3.83	3.33	3.88	3.27	3.45	3.43	3.76	3.35

This table reports the OLS regression on earnings management (EM). The sample includes 75 U.S.-listed Chinese companies during 2002–2014.The total number of observations is 449. The sample is obtained from the S&P Capital IQ and supplemented from the Wind Database. LP is a dummy for listing place, FN is a dummy for state-owned enterprise, AOT is a dummy for auditor going concern, and BIG4is a dummy for Big 4 auditing firms. LEV is the ratio of total debt to total assets. LNSIZE is the logarithm of firm's total assets. CH is the ratio of cash to total assets. REV is the change in sales. CF is the ratio of cash from operating to total assets. ROA is return on assets. Positive, negative, uncertain, litigious, strong modal, moderate modal, weak modal, and irregular verbs are the eight categories of wordlist defined from the 2014 Master Dictionary of McDonald, and Loughran and McDonald (2011) word list. All regressions include firm, industry and year fixed effects. The t-statistics are in parentheses

***, **, and* denote significance at the 1%, 5%, and 10% level respectively

Uncertainty, *Strongml*, *Moderml*, *and Weakml*) have positive and statistically significant coefficients for *E.M*. In Column (2), *Positive* has a positive and significant coefficient (6.482), suggesting that firms using a higher proportion of positive words in their 20-Fs experience more earnings management. This finding implies that companies use positive words more frequently in financial statements to conceal their true earnings status. Similarly, *uncertainty* has a positive coefficient (3.793), suggesting that firms using higher proportions of uncertain words in their 20-Fs engage in more earnings management. A one standard deviation increase in the percentage of *uncertainty* leads to a 0.009 increase in *EM* (3.793 multiplied by the standard deviation of 0.00234). This suggests that companies that frequently use words indicating uncertainty about the future may engage in more earnings management activities than others. Similarly, Columns (6)–(8) show that *Strongml*, *Moderml*, and *Weakml* have positive and significant coefficients (6.372, 8.741, and 5.357, respectively), suggesting that firms that use a high proportion of modal words engage in more earnings management. Specifically, a one standard deviation change in the percentage of *Strongml*, *Moderml*, and *Weakml* is linked to 0.005 (6.372 times 0.0008), 0.005 (8.741 times 0.00058), and 0.008 (5.357 times 0.00153) increase in earnings management, respectively.[5] We also find that the coefficients of *REV* and *ROA* are positive and significant. A possible explanation is that companies with better growth prospects or higher profitability ratios may have more motives for earnings management and manipulating numbers (Espahbodi et al. 2021). Overall, the results demonstrate that the tone of 20-Fs relates to earnings management. Specifically, a positive correlation exists between a positive, uncertain, or modal tone in 20-Fs and earnings management behaviors, suggesting that the sentiment words "good, may, could, depend, and approximately" may reflect managers' immoral behaviors. In summary, the results support both hypotheses and indicate that qualitative textual tones can provide valuable information to investors.

These results are consistent with those obtained for the U.S. companies in previous studies (Jegadeesh and Wu 2013; Kang et al. 2018; Jaspersen et al. 2021). For example, Kang et al. (2018) claim that it is risky to interpret business managers' positive expressions because the data can include their subjective opinions and viewpoints from the companies' perspectives. Furthermore, they employ the text-mining method to identify the tones of the 10-K narratives to determine whether the changes are consistent with the current earning levels. Jaspersen et al. (2021) suggest that qualitative disclosures are additional sources of information about a company's financial situation. Still, executives likely hide their earnings management activities in these disclosures. Nevertheless, their results demonstrate that qualitative disclosures can predict earnings management and are useful for learning about companies' accounting choices. In addition, Jegadeesh and Wu (2013) find a significant relationship between document tone and market reaction to 10-K filings for negative and positive words. Furthermore, they indicate a need to partition words into positive and negative word lists subjectively. However, our study focuses on 20-Fs, different from previous studies on 10-Ks.

[5] It further corrects the standard errors for multiple hypothesis testing, which can test the significance of the eight main sentiment variables. Following Jaspersen et al. (2020), this study uses the Bonferroni correction method and finds that the main regression results are robust. For brevity, it does not report the results in the manuscript, but the results are available upon request.

Table 4 Multicollinearity test

Variables	Positive	Negative	Uncertainty	Litigious	Strongml	Moderml	Weakml	Irrverb
Independent variables: Sentiment words								
LNSIZE	3.69	3.59	3.65	3.58	3.73	3.84	3.70	3.61
CF	2.64	2.63	2.64	2.64	2.63	2.63	2.64	2.63
ROA	2.54	2.52	2.53	2.53	2.53	2.53	2.53	2.51
FN	2.06	2.03	2.07	2.03	2.28	2.16	2.01	2.35
CH	1.93	2.01	1.94	2.02	1.95	1.92	1.94	1.97
LP	1.64	1.64	1.64	1.64	1.64	1.64	1.65	1.65
LEV	1.51	1.52	1.53	1.52	1.51	1.52	1.52	1.52
BIG4	1.48	1.50	1.48	1.50	1.49	1.48	1.48	1.50
Sentiment words	1.14	1.19	1.15	1.18	1.21	1.23	1.21	1.28
REV	1.13	1.14	1.12	1.16	1.13	1.13	1.13	1.14
AOT	1.06	1.06	1.06	1.06	1.06	1.06	1.06	1.06
Mean VIF	1.89	1.89	1.89	1.89	1.92	1.92	1.90	1.93

This table reports the mean VIF values of selected variables. The sample includes 75 U.S.-listed Chinese companies during 2002–2014.The total number of observations is 449. The sample is obtained from the S&P Capital IQ and supplemented from the Wind Database. LP is a dummy for listing place, FN is a dummy for state-owned enterprise, AOT is a dummy for auditor going concern, and BIG4is a dummy for Big 4 auditing firms. LEV is the ratio of total debt to total assets. LNSIZE is the logarithm of firm's total assets. CH is the ratio of cash to total assets. REV is the change in sales. CF is the ratio of cash from operating to total assets. ROA is return on assets. Positive, negative, uncertain, litigious, strong modal, moderate modal, weak modal, and irregular verbs are the eight categories of wordlist defined from the 2014 Master Dictionary of McDonald, and Loughran and McDonald (2011) word list

Multicollinearity test

A possible problem in the baseline regression models is the strong correlation between the variables and relatively small sample size, leading to multicollinearity. This study conducts an extra multicollinearity test commonly used in multiple regression models and reports the variance inflation factors (VIFs) in Table 4 to address this issue. It shows that the average VIF values are less than 10 for all multiple regressions, suggesting that there are no significant multicollinearity issues between the sentiment variables. These results confirm the robustness of the baseline regression results.

An alternative earnings management variable

This study further investigates whether the impact of textual analysis on earnings management persists when an alternative measure of earnings management is used to check the robustness of the empirical results. The performance-adjusted discretionary accruals variable developed by Kothari et al. (2005) is the alternative measure. It estimates the following regression in Eq. (3):

$$TAccr_{it} = \alpha_0 + \alpha_1 \left(\frac{1}{Asset_{i,t-1}} \right) + \alpha_2 \Delta Rev_{it} + \alpha_3 PPE_{it} + \alpha_4 ROA_{it} + \varepsilon_{it} \tag{3}$$

where $TAccr_{it}$ is total accruals, measured as the change in non-cash current assets minus the change in current non-interest-bearing liabilities, minus depreciation and amortization expenses for firm i in year t, scaled by lagged total assets ($Asset_{i,t-1}$);ΔRev_{it} is the annual change in revenue scaled by lagged total assets; PPE_{it} is property, plant, and equipment costs for firm i in year t, scaled by lagged total assets; and ROA_{it} is the return on investments for firm i in year t. The residuals from the regression model are discretionary accruals. This study uses the absolute value of discretionary accruals ($AVDA$) as

Table 5 Regressions using an alternative measure of earnings management

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent variable: AVDA									
Positive		9.097***							
		(2.86)							
Negative			1.603						
			(1.14)						
Uncertainty				4.392**					
				(2.44)					
Litigious					0.404				
					(0.46)				
Strongml						6.183			
						(1.16)			
Moderml							14.935**		
							(1.98)		
Weakml								6.320**	
								(2.25)	
Irrverb									3.560
									(0.97)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	449	449	449	449	449	449	449	449	449
Adj R^2	0.337	0.349	0.337	0.344	0.335	0.337	0.341	0.343	0.336
F- statistics	8.49	8.62	8.26	8.51	8.21	8.27	8.40	8.46	8.25

This table reports the OLS regression results using an alternative measure of earnings management (AVDA). The dependent variable is the estimates of discretionary revenues developed by Kothari et al. (2005). The sample includes 75 U.S.-listed Chinese companies during 2002–2014.The total number of observations is 449. The sample is obtained from the S&P Capital IQ and supplemented from the Wind Database. Positive, negative, uncertain, litigious, strong modal, moderate modal, weak modal, and irregular verbs are the eight categories of wordlist defined from the 2014 Master Dictionary of McDonald, and Loughran and McDonald (2011) word list. For the sake of brevity, we use *Controls* to represent all the control variables. All regressions include firm, industry, and year fixed effects. The t-statistics are in parentheses

***, **, and* denote significance at the 1%, 5%, and 10% level, respectively

an alternative proxy for earnings management. Table 5 shows the regression results. As shown in Columns (2), (4), (7), and (8) of Table 5, the significant relationship between most sentiment words and earnings management continues to hold as an alternative measure of earnings management. This result confirms the main results presented in Table 3. This shows that companies that use more positive or uncertain words in their financial reports are more likely to engage in earnings management.

Controlling for the length of financial statements

The percentage of sentiment words used in financial statements may be affected by the text length. In this study, we use file size, measured by the natural logarithm of the file size in megabytes (Lnfsize) of the "complete submission text file" for Form 20-F filing, as a proxy for the length of the text. As shown in Table 6, the significance levels and signs of the coefficients of the sentiment word variables do not change significantly, indicating that the baseline regression results are robust after controlling for the length of the financial statements.

Table 6 Regressions after controlling for the length of financial statement

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent variable: EM								
Positive	7.401***							
	(3.36)							
Negative		0.432						
		(0.47)						
Uncertainty			4.688***					
			(3.66)					
Litigious				0.231				
				(0.48)				
Strongml					4.826			
					(1.44)			
Moderml						6.586**		
						(1.26)		
Weakml							6.170***	
							(3.10)	
Irrverb								1.732
								(0.76)
Lnfsize	0.002	−0.004	0.004	−0.005*	−0.003	−0.002	0.002	−0.004
	(0.67)	(−1.28)	(1.02)	(−1.76)	(−0.84)	(−0.66)	(0.62)	(−1.23)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	448	448	448	448	448	448	448	448
Adj R^2	0.163	0.141	0.167	0.141	0.145	0.144	0.160	1.142
F- statistics	3.72	3.29	3.81	3.29	3.36	3.34	3.66	3.30

This table reports the OLS regression results after controlling for the length of financial statement (Lnfsize). The dependent variable is the earnings management (EM). The sample includes 75 U.S.-listed Chinese companies during 2002–2014.The total number of observations is 449. The sample is obtained from the S&P Capital IQ and supplemented from the Wind Database. Positive, negative, uncertain, litigious, strong modal, moderate modal, weak modal, and irregular verbs are the eight categories of wordlist defined from the 2014 Master Dictionary of McDonald and Loughran and McDonald (2011) word list. For the sake of brevity, we use *Controls* to represent the control variables except for Lnfsize. All regressions include firm, industry, and year fixed effects. The t-statistics are in parentheses

***, **, and* denote significance at the 1%, 5%, and 10% level, respectively

Impact of the 2008 financial crisis

This study also analyzes the impact of the financial crisis on earnings management because firms may face different regulations during crisis periods, influencing how they report earnings. Additionally, firms may change the tone of their financial disclosures during a crisis period because of increased uncertainty. Therefore, it divides the sample period into two subsamples: 2002–2007 (pre-crisis) and 2008–2014 (post-crisis). This analysis identifies whether there is a difference in the tone of financial statements before and after the crisis.

The results for two subsamples are reported in Panels A and B in Table 7. The first column of each panel includes only control variables. Columns (2)–(9) add the variables for each sentiment category as the main independent variables. *Positive* has positive

Table 7 The impact of financial crisis on earnings management

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: before financial crisis 2002–2007									
Positive		10.10**							
		(2.03)							
Negative			−0.426						
			(−0.23)						
Uncertainty				1.280					
				(0.56)					
Litigious					−0.682				
					(−1.11)				
Strongml						4.046			
						(0.64)			
Moderml							−8.134		
							(−0.81)		
Weakml								0.144	
								(0.04)	
Irrverb									−5.993*
									(−1.94)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year`	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	86	86	86	86	86	86	86	86	86
Adj R^2	0.357	0.387	0.347	0.350	0.359	0.351	0.353	0.347	0.384
Panel B: after financial crisis 2008–2014									
Positive		6.991***							
		(3.61)							
Negative			1.270						
			(1.46)						
Uncertainty				4.324***					
				(3.93)					
Litigious					0.622				
					(1.06)				
Strongml						7.533**			
						(3.36)			
Moderml							10.170**		
							(2.19)		
Weakml								6.296***	
								(3.64)	
Irrverb									4.603*
									(1.95)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	363	363	363	363	363	363	363	363	363
Adj R^2	0.112	0.142	0.114	0.148	0.112	0.122	0.121	0.143	0.119

This table reports the impact of financial crisis on earnings management. The sample is obtained from the S&P Capital IQ and supplemented from the Wind Database. Positive, negative, uncertain, litigious, strong modal, moderate modal, weak modal, and irregular verbs are the eight categories of wordlist defined from the 2014 Master Dictionary of McDonald and Loughran and McDonald (2011) word list. For the sake of brevity, we use *Controls* to represent all the control variables. Regressions results are reported for two sub-periods: 2002–2007 and 2008–2014. The dependent variable is the estimates of discretionary revenues, proposed by McNichols and Stubben (2008) and Stubben (2010), as a proxy of earnings

Table 7 (continued)

management. There are 86 observations during 2002–2007 and 363 observations during 2008–2014. All regressions include firm, industry, and year fixed effects. The t-statistics are in parentheses

***, **, and * denote significance at the 1%, 5%, and 10% level, respectively

coefficients on earnings management for both periods (10.10 and 6.991, respectively), suggesting no significant difference in the relationship between positive words and earnings management in these periods. Column (9) of Panel A shows that *Irrverb* has a marginally significant negative coefficient of − 5.993. However, the results may be biased due to the small sample size for this category (only 86 observations). Therefore, *Irrverb* has less power to explain earnings management in the pre-crisis subsample. Columns (3)–(9) of Panel B in Table 7 show that the sentiment variables *Uncertainty*, *Strongml*, *Moderml*, and *Weakml* have significant positive coefficients on *EM* after the crisis, consistent with the results for the full sample in Table 3. The results confirm that these modal variables have similar and significant relationships with *EM* in the pre -and post-crisis periods.

Overall, there is no significant change in the relationship between *Positive*, *Uncertainty*, *Strongml*, *Moderml*, and *Weakml* and *EM* in the pre -and post-crisis periods. Therefore, the main results are robust after controlling for the effects of the 2008 financial crisis.

Information technology versus non-information technology industries

The sample reveals that 42.98% of observations are from the information technology industry. Thus, we divide the sample into two subsamples: information technology and non-information technology industries. Panels A and B of Table 8 present the regression results for the information technology industry and non-information technology industry subsamples, respectively. Columns (4)–(6) of Panel A show that the coefficients of *Litigious*, *Strongml*, and *Moderml* are significant, indicating that firms in information technology are more likely to use a positive tone in their financial disclosures to conceal earnings management. However, there are no significant changes in the signs of the estimated coefficients, indicating that the robustness of the empirical results holds for all industries.

Discussion of the empirical results

The main results are summarized as follows. First, positive, uncertain, and modal words are positively and significantly correlated with earnings management. Table 3 shows that firms with a higher proportion of positive, uncertain, and modal words in financial reports are more likely to engage in earnings management. The results imply that companies use more positive words when concealing earnings management, probably to attract potential investors or other businesses. Second, although the VIF values indicated possible multicollinearity among the selected variables, the mean values of the VIFs are less than 10 in each regression. This suggested no significant multicollinearity issues in the main regression models. Third, it uses the absolute values of discretionary accruals as an alternative proxy for earnings management and re-estimates the main regression. These results provide further support for the main results in Table 3, even after controlling for the effect of the length of financial statements. In addition, this study considers the impact of the financial crisis and finds that the main results hold in the pre- and

Table 8 Information technology industry versus non-information technology industry

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A (Information technology industry)								
Positive	9.877***							
	(3.09)							
Negative		1.379						
		(1.13)						
Uncertainty			4.664***					
			(2.83)					
Litigious				1.292*				
				(1.72)				
Strongml					11.25**			
					(2.07)			
Moderml						17.52**		
						(3.39)		
Weakml							5.511**	
							(2.14)	
Irrverb								4.381
								(1.33)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	193	193	193	193	193	193	193	193
Adj. R^2	0.126	0.020	0.003	0.013	0.009	0.006	0.008	0.017
Panel B (Non-information technology industry)								
Positive	4.459**							
	(2.25)							
Negative		0.189						
		(0.19)						
Uncertainty			2.414**					
			(2.01)					
Litigious				−0.686				
				(−1.12)				
Strongml					3.466			
					(1.07)			
Moderml						2.919		
						(0.62)		
Weakml							4.162**	
							(2.12)	
Irrverb								1.540
								(0.63)
Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	256	256	256	256	256	256	256	256
Adj. R^2	0.241	0.224	0.238	0.228	0.228	0.225	0.239	0.225

This table reports the regression results for both the information technology industry and non-information technology industry. The sample is obtained from the S&P Capital IQ and supplemented from the Wind Database. Positive, negative, uncertain, litigious, strong modal, moderate modal, weak modal, and irregular verbs are the eight categories of wordlist defined from the 2014 Master Dictionary of McDonald and Loughran and McDonald (2011) word list. For the sake of brevity, we use *Controls* to represent all the control variables. The dependent variable is the estimates of discretionary revenues, proposed by McNichols and Stubben (2008) and Stubben (2010), as a proxy of earnings management. There are 193

Table 8 (continued)

observations for information technology industry and 256 for non-information technology industry. For the sake of brevity, we use *Controls* as all the controlled variables. All regressions include firm, industry, and year fixed effects. The t-statistics are in parentheses

***, **, and * denote significance at the 1%, 5%, and 10% level, respectively

post-crisis periods. Finally, it divides the sample into information technology and non-information technology industries and finds that the results do not exhibit significant differences between these two sub-samples.

Conclusions

This study uses a sample of 153 U.S.-listed Chinese firms to investigate the relationship between earnings management behavior and the tone of 20-F filings. We find that firms that employ more positive words in their 20-Fs engage in more earnings management, as concealing the firms' true earnings status may help attract investors. Specifically, firms using more positive, uncertain, or modal words in their 20-Fs engage in more earnings management behaviors, suggesting that textual tone is informative. Textual analysis can be a valuable tool for investors seeking to identify earnings management. The main results are robust after conducting a multicollinearity test, re-estimating the regression models using an alternative proxy for earnings management, considering the effect of the financial crisis, and testing the results separately for information and non-information technology industries.

This study provides an innovative textual analysis method that uses financial information from Form 20-F to identify earnings management in the Chinese context. Li et al. (2021b) claim that owing to the complexity of human behaviors and changing social environments, the distributions of financial data are usually complex, and it is challenging to find patterns and provide reasonable interpretations. This study extends the literature using Loughran and McDonald's word lists as earnings management indicators. The results confirm that the tone in 20-Fs correlates with earnings management behaviors; specifically, a higher proportion of positive, uncertain, or modal words like "good, may, could, depend, and approximately" in the 20-Fs implies more earnings management behaviors. The results are consistent with impression management theory that the strategic positioning of positive and negative words within a chief executive officer (CEO) letter is a subtle form of impression management. Additionally, managers present information in such an order that the reader of the CEO letter has a more positive perception of the underlying message (Boudt and Thewissen 2019). Moreover, Li (2008) finds that the annual reports of firms with lower earnings are harder to read and that firms with easier-to-read annual reports have more persistent positive earnings. This study has important implications for international investors, analysts, and legislators to adopt an effective tool for identifying earnings management behaviors in financial statements.

This study has some limitations. For example, the Loughran and McDonald list (2011) is a six-category list later modified to contain eight categories in our study. This method is based on the unique words used in financial reports. Thus, a better theoretical framework is needed to explain the association between tone categories and earnings management. In addition, 2014 was the end date for the sample. Therefore, extending the sample to more recent years is possible, but we only consider the years from the pre- and

post-2008 samples. Furthermore, the main measure of earnings management is a simplified version of the traditional short-term Jones Model. Although we use a modified version of the Jones Model proposed by Kothari et al. (2005), problems associated with the main measure may still be present because they are derived from the same model. Future studies should include alternative measures not based on the Jones Model.

Appendix 1: Variable definitions

Variables	
EM	Earnings management, the absolute values of discretionary revenues, a proxy is estimated by Eq. (1)
Positive	It is calculated by the percentage of words in the 20-F that are classified as positive using the 2014 Master Dictionary of McDonald, and Loughran and McDonald (2011) word list. They are some positive ingredients expressing good, inspirational, motivational, and encouraging meanings. Examples include *beneficial, successful, good, achieved, and empower,* etc
Negative	It is calculated by the percentage of words in the 20-F that are classified as negative using the 2014 Master Dictionary of McDonald, and Loughran and McDonald (2011) word list. They are marked by denial, prohibition, or refusal or by absence, withholding, or removal of something positive. Examples include *loss, failure, abandon, and decline,* etc
Uncertainty	It is calculated by the percentage of words in the 20-F that are classified as uncertain using the 2014 Master Dictionary of McDonald, and Loughran and McDonald (2011) word list. They are some words giving unclear conclusions or suggestions. Examples include *almost, and nearly,* etc
Litigious	It is calculated by the percentage of litigious words in the 20-F using the 2014 Master Dictionary of McDonald, and Loughran and McDonald (2011) word list. They are some words relating to, or characterized by litigation. Examples include *abovementioned, abrogated, and certiorari*, etc
Strongml	It is calculated by the percentage of strong modal words in the 20-F using the 2014 Master Dictionary of McDonald, and Loughran and McDonald (2011) word list. They are some words expressing a necessity. Examples include *must, never, definitely, and will,* etc
Moderml	It is calculated by the percentage of moderate modal words in the 20-F using the 2014 Master Dictionary of McDonald, and Loughran and McDonald (2011) word list. They are some words expressing the degree of modality of people's opinions or argument. Examples include *can, generally, and usually,* etc
Weakml	It is calculated by the percentage of weak modal words in the 20-F using the 2014 Master Dictionary of McDonald, and Loughran and McDonald (2011) word list. They are some words expressing a possibility. Examples include *could, should, and ought to,* etc
Irrverb	It is calculated by the percentage of irregular verbs in the 20-F using the 2014 Master Dictionary of McDonald, and Loughran and McDonald (2011) word list. They are verbs that follow a different pattern. Examples include *beat, cut, and forgot,* etc
LP	A dummy variable to control for listing place, it equals 1 if the firm is listed on the NASDAQ, and 0 otherwise
FN	Firm characteristic as 1 if the controlling shareholder is state-owned, and 0 otherwise
AOT	Auditor going concern, it equals 1 if auditor going concern doubts, and 0 otherwise
BIG4	Auditing firm characteristic, it equals 1if company is audited by the Big 4 auditing firms, and 0 otherwise
LEV	The ratio of total debt to total assets
LNSIZE	The natural logarithm of firm's total assets
CH	The ratio of cash and cash equivalents to total assets
REV	The change in sales of the company
CF	The ratio of cash from operating to total assets
ROA	Return on assets

Appendix 2: Sample description by industry

GSIC	Industry	Frequency	Proportion (%)
10	Energy	36	8.02
15	Materials	37	8.24
20	Industrials	60	13.36
25	Consumer discretionary	75	16.70
30	Consumer staples	11	2.45
35	Health care	17	3.79
45	Information technology	193	42.98
50	Communication services	12	2.67
55	Utilities	8	1.78
	Total	449	100.00

Abbreviations
AMEX: American stock exchange; CEO: Chief executive officer; CSMAR: China stock market and accounting research database; EDGAR: Electronic data gathering, analysis, and retrieval; FCPA: Foreign Corrupt Practices Act; GICS: Global industry classification standard; HFCAA: Holding Foreign Companies Accountable Act; IPO: Initial public offering; MD&A: Management discussion and analysis; NASDAQ: National association of securities dealers automated quotations; NYSE: New York stock exchange; OLS: Ordinary least square; SEC: Securities and exchange commission.

Acknowledgements
The authors would like to thank the participants of the Southwestern Finance Association (SWFA) 2019 for their helpful comments and suggestions. We are grateful to Dr. Guangjun Cao for his helpful comments and contribution to this manuscript.

Authors' contributions
All authors read and approved the final manuscript.

Funding
This research is supported by the National Social Science Foundation of China (Grant No.17BJY019).

Availability of data and materials
The data and materials are available upon request.

Declarations

Ethical approval and consent to participate
This article does not contain any studies with human participants or animals performed by any of the authors.

Consent for publication
We confirm that all authors have read and approved the manuscript for submission. We also confirm that the content of the manuscript has not been published or submitted for publication elsewhere. All authors contributed equally to this study. We are solely responsible for any errors and omissions.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Economics and Finance, Xi'an Jiaotong University, Xi'an, China. [2]Hanwang Technology Co., Ltd., Beijing, China. [3]Archie W. Dunham College of Business at Houston Baptist University, Houston, USA.

Received: 8 June 2021 Accepted: 16 March 2022
Published online: 10 May 2022

References

Abbasi A, Chen H, Salem A (2008) Sentiment analysis in multiple languages: feature selection for opinion classification in Web forums. ACM Trans Inf Syst 26:1–34
Ang JS, Jiang ZQ, Wu CP (2016) Good apples, bad apples: sorting among Chinese companies traded in the U.S. J Bus Ethics 134(4):611–629
Bao SR, Lewellyn KB (2017) Ownership structure and earnings management in emerging markets—an institutionalized agency perspective. Int Bus Rev 26(5):828–838
Beckmann KS, Escobari DA, Ngo T (2019) The real earnings management of cross-listing firms. Glob Finance J 41:128–145

Bian SB, Jia DK, Li RH, Sun WJ, Yan ZP, Zheng YF (2021) Can management tone predict IPO performance? Evidence from mandatory online roadshows in China. Pac Basin Finance J 68:101588
Blau BM, DeLisle JR, Price SM (2015) Do sophisticated investors interpret earnings conference call tone differently than investors at large? Evidence from short sales. J Corp Finance 31:203–219
Bodnaruk A, Loughran T, McDonald B (2015) Using 10-K text to gauge financial constraints. J Financ Quant Anal 50(4):1–24
Boudt K, Thewissen J (2019) Jockeying for position in CEO letters: impression management and sentiment analytics. Financ Manag 48(1):77–115
Chen YJ (2014) Detecting Fraud in Narrative Annual Reports. Available at https://doi.org/10.2139/ssrn.2511629
Chen XQ, Li WP, Chen ZF, Huang JS (2021) Environmental regulation and real earnings management—evidence from the SO_2 emissions trading system in China. Finance Res Lett. https://doi.org/10.1016/j.frl.2021.102418
Demaline CJ (2019) Disclosure characteristics of firms being investigated by the SEC. J Corp Account Finance 30(4):11–24
Dong NY, Wang FJ, Zhang JR, Zhou J (2020) Ownership structure and real earnings management: evidence from China. J Account Public Policy 39(3):106733
Dou Y, Khan M, Zou Y (2016) Labor unemployment insurance and earnings management. J Account Econ 61:166–184
El Diri M, Lambrinoudakis C, Alhadab M (2020) Corporate governance and earnings management in concentrated markets. J Bus Res 108:291–306
Espahbodi R, Liu N, Weigand RA (2021) Opportunistic earnings management or performance-related effects? Evidence from dividend-paying firms. Glob Finance J. https://doi.org/10.1016/j.gfj.2021.100636
Fama EF, French KR (1992) The cross-section of expected stock returns. J Finance 47(2):427–465
Fisher IE, Garney MR, Hughes ME (2016) Natural language processing in accounting, auditing, and finance: a synthesis of the literature with a roadmap for future research. Intell Syst Account Finance Manag 23(3):157–214
Frankel R, Mayew WJ, Sun Y (2010) Do pennies matter? Investor relations consequences of small negative earnings surprises. Rev Account Stud 15(1):220–242
Gandía JL, Huguet D (2021) Textual analysis and sentiment analysis in accounting. Rev Contab Span Account Rev 24(2):2021
García Lara JM, García Osma B, Penalva F (2020) Conditional conservatism and the limits to earnings management. J Account Public Policy 39(4):106738
Henry E (2008) Are investors influenced by how earnings press releases are written? J Bus Commun 45(4):363–407
Hofstede GJ, Hofstede M (2010) Cultures and Organizations: software of the mind, 3rd edn. McGraw-Hill, London
Hu JC (2021) Do facilitation payments affect earnings management? Evidence from China. J Corp Finance 68:101936. https://doi.org/10.1016/j.jcorpfin.2021.101936
Huang X, Teoh SH, Zhang Y (2014) Tone management. Account Rev 89(3):1083–1113
Jaeschke R, Lopatta K, Yi C (2018) Managers' use of language in corrupt firms' financial disclosures: evidence from FCPA violators. Scand J Manag 34(2):170–192
Jaspersen JG, Ragin MA, Sydnor JR (2020) Linking subjective and incentivized risk attitudes: the importance of losses. J Risk Uncertain 60(2):187–206
Jaspersen JG, Richter A, and Zoller S (2021) Predicting earnings management from qualitative disclosures. Munich Risk and Insurance Center Working Paper 40. https://doi.org/10.2139/ssrn.3732203
Jegadeesh N, Wu D (2013) Word power: a new approach for content analysis. J Financ Econ 110(3):712–729
Jiang F, Lee J, Martin X, Zhou G (2019) Manager sentiment and stock returns. J Financ Econ 132(1):126–149
Kang T, Park DH, Han I (2018) Beyond the numbers: the effect of 10-K tone on firms' performance predictions using text analytics. Telemat Inform 35(2):370–381
Kearney C, Liu S (2014) Textual sentiment in finance: a survey of methods and models. Int Rev Financ Anal 33:171–185
Khokhar AR (2011) Firm size, information asymmetry and window dressing in cash holdings: evidence from quarterly financial statements. Midwest Finance Association 2012 annual meetings paper, Available at SSRN: https://ssrn.com/abstract=1865361 or https://doi.org/10.2139/ssrn.1865361
Kim J, Kim Y, Zhou J (2017) Languages and earnings management. J Account Econ 63(2–3):288–306
Kothari SP, Leone AJ, Wasley CE (2005) Performance matched discretionary accrual measures. J Account Econ 39(1):163–197
Lang M, Stice-Lawrence L (2015) Textual analysis and international financial reporting: large sample evidence. J Account Econ 60(2–3):110–135
Li F (2008) Annual report readability, current earnings, and earnings persistence. J Account Econ 45(2–3):221–247
Li T, Zaitas N (2017) Information environment and earnings management of dual class firms around the world. J Bank Finance 74:1–23
Li T, Kou G, Peng Y, Yu PS (2021a) An integrated cluster detection, optimization, and interpretation approach for financial data. IEEE Trans Cybern. https://doi.org/10.1109/TCYB.2021.3109066
Li K, Mai F, Shen R, Yan X (2021b) Measuring corporate culture using machine learning. Rev Financ Stud 34(7):3265–3315
Loughran T, McDonald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. J Finance 66(1):35–65
Loughran T, McDonald B (2013) IPO first-day returns, offer price revisions, volatility, and form S-1 language. J Financ Econ 109(2):307–326
Loughran T, Mcdonald B (2014) Measuring readability in financial disclosures. J Finance 69(4):1643–1671
Loughran T, Mcdonald B (2016) Textual analysis in accounting and finance: a survey. J Account Res 54(4):1187–1230
Luo Y, Fang F, Esqueda O (2012) The overseas listing puzzle: post-IPO performance of Chinese stocks and ADRs in the U.S. market. J Multinatl Financ Manag 22(5):193–211
Mai F, Tian S, Lee C, Ma L (2019) Deep learning models for bankruptcy prediction using textual disclosures. Eur J Oper Res 274(2):743–758
McNichols MF, Stubben SR (2008) Does earnings management affect firms' investment decisions? Account Rev 83(6):1571–1603

Ng J, Wu H, Zhai WH, Zhao J (2021) The effect of shareholder activism on earnings management: evidence from shareholder proposals. J Corp Finance 69:102014
Ni XR (2020) Does stakeholder orientation matter for earnings management: evidence from non-shareholder constituency statutes. J Corp Finance 62:101606
Pejic Bach M, Krstic Z, Seljan S (2019) Big data text mining in the financial sector. In: Metawa N, Elhoseny M, Hassanien A, Hassan M (eds) Expert systems in finance: smart financial applications in big data environments. Routledge, London, pp 80–96
Pevzner M, Xie F, Xin X (2015) When firms talk, do investors listen? The role of trust in stock market reactions to corporate earnings announcements. J Financ Econ 117(1):190–223
Price MK, Doran JS, Peterson DR, Bliss A (2012) Earnings conference calls and stock returns: the incremental informativeness of textual tone. J Bank Finance 36(4):992–1011
Shan YG (2019) Do corporate governance and disclosure tone drive voluntary disclosure of related-party transactions in China? J Int Account Audit Tax 34:30–48
Sims T, Tsai JL, Jiang D, Wang Y, Fung HH, Zhang X (2015) Wanting to maximize the positive and minimize the negative: implications for mixed affective experience in American and Chinese contexts. J Personal Soc Psychol 109(2):292–315
Stubben S (2010) Discretionary revenues as a measure of earnings management. Account Rev 85(2):695–717
Uzkurt C, Kumar R, Semih Kimzan H, Eminoğlu G (2013) Role of innovation in the relationship between organizational culture and firm performance: a study of the banking sector in Turkey. Eur J Innov Manag 16(1):92–117
Wei L, Li G, Zhu X, Sun X, Li J (2019) Developing a hierarchical system for energy corporate risk factors based on textual risk disclosures. Energy Econ 80:452–460
Wu DX, Yao X, Guo JL (2021) Is textual tone informative or inflated for firm's future value? Evidence from Chinese listed firms. Econ Model 94:513–525
Yang X, Luo Y (2014) Rumor clarification and stock returns: do bull markets behave differently from bear markets? Emerg Mark Finance Trade 50(1):197–209
Zha Q, Kou G, Zhang H et al (2020) Opinion dynamics in finance and business: a literature review and research opportunities. Financ Innov 6:44

Publisher's Note