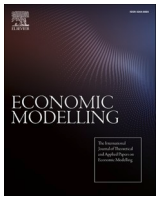
Economic Modelling 126 (2023) 106428

Contents lists available at ScienceDirect

Economic Modelling

journal homepage: www.journals.elsevier.com/economic-modelling



Textual analysis and detection of financial fraud: Evidence from Chinese manufacturing firms

Jing Li, Nan Li, Tongshui Xia *, Jinjin Guo

Business School, Shandong Normal University, 1 Daxue Rd, Changqing, Jinan, Shandong Province, 250000, China

ARTICLE INFO

Keywords:
textual analysis
Financial fraud
Language structure
Language quality
Language expression

ABSTRACT

With the increasing complexity of financial statement manipulation, relying solely on quantitative financial data may not effectively detect financial fraud. While textual analysis can provide additional insight, little research has been conducted on its multiple dimensions. Using 579 listed Chinese manufacturing firms in 2020, we select readability, forward-looking, similarity, matching degree, and positive and negative sentiment indicators from textual language structure, quality, and expression of management discussion and analysis texts, in combination with financial indicators, to detect financial fraud. Our findings indicate that fraudulent firms tend to be overly cautious in their financial reporting, express fewer positive sentiments, and conceal financial fraud by increasing the complexity of their annual reports and using more degree adverbs to modify forward-looking information. This study also highlights the importance of considering textual language expression in detecting financial fraud in state-owned and non-state-owned firms.

1. Introduction

Financial fraud is a global phenomenon that causes substantial losses for investors and undermines the efficient allocation of resources and the functioning of securities markets (Kong et al., 2021). The Enron scandal in 2001 had a long-lasting and far-reaching impact on capital markets in the United States and overseas (Gillan and Martin, 2007), while the associated auditing failures forced the company's accounting firm—Arthur Andersen—to dissolve. More recently, General Electric (GE) was found to have misled investors through a series of disclosure failures in 2019 amounting to US$38 billion, equivalent to more than 40% of GE's market value. In China, Kangmei Pharmaceuticals' financial fraud in 2020 totaled 30 billion yuan, the largest financial fraud case in the history of the A-shares market. Prior to the investigation, the firm's market value decreased by nearly 90%, resulting in a significant loss for investors.

The China Securities Regulatory Commission (CSRC) has encountered increasingly complex and covert methods of fraud, which often come to light after significant losses have occurred (Dyck et al., 2010). Detecting financial fraud has become challenging, especially in the context of the COVID-19 pandemic, which has created systemic financial risks and raised concerns about a global economic slowdown (Zhu et al., 2021). Firms may be more likely to engage in financial fraud due to incentives from opportunistic venture capitalists (Que and Zhang, 2019). Detecting financial fraud remains an important and challenging task for accountants, auditors, and regulators. However, analyzing structured financial data is tedious and rigid, while financial fraud tactics are becoming sophisticated. In addition, methods of whitewashing financial data have matured, and analytical procedures that use financial data only are likely to be ineffective in detecting fraud (Zhang et al., 2022). Thus, users of reports should also consider the supplementary information included in the accounting reports, such as text.

Researchers have increasingly recognized the significance of the management discussion and analysis (MD&A) section in annual reports as a valuable source of information. MD&A provides a comprehensive evaluation of a firm's past performance and future prospects, complementing the financial data (Maharjan and Lee, 2021). Craja et al. (2020) employ a hierarchical attention network to extract text features from the MD&A sections of annual reports, capturing the content and context of managerial comments, which serve as supplementary predictors for detecting fraud.

Existing research focuses on the language characteristics of text, specifically its intonation (Durnev and Mangen, 2020) and readability (Wang L. et al., 2021). Although previous studies have contributed to the detection of financial fraud using textual information, they have often focused on only one aspect of report information without further

* Corresponding author.
E-mail addresses: lijing5329@163.com (J. Li), linan987@sdnu.edu.cn (N. Li), 117011@sdnu.edu.cn (T. Xia), 78493286@qq.com (J. Guo).

https://doi.org/10.1016/j.econmod.2023.106428
Received 17 July 2022; Received in revised form 25 June 2023; Accepted 26 June 2023
Available online 1 July 2023

refining the indicators from the perspective of language classification. For a more accurate detection of financial fraud, multiple dimensions of language characteristics can be considered.

In this study, we select three language dimensions from the MD&A sections of financial reports, in combination with financial indicators, to detect financial fraud. Additionally, we investigate the mechanism behind the significant text indicators in identifying financial fraud. We contribute to literature in four ways. First, we refine three dimensions from language characteristics to construct a comprehensive framework of MD&A text indicators—structure, quality, and expression, which expand the sources of MD&A text indicators beyond previous studies that focus solely on readability or intonation. Second, incorporating textual indicators disclosed in the MD&A section with financial indicators improves the accuracy of financial fraud detection. Third, we test the plausibility of the results using various approaches, including the logistic regression, extreme gradient boosting (XGBoost), and multilayer perceptron neural network (MLP NN) models, with the XGBoost model having the highest recognition accuracy. Finally, this study emphasizes the significance of textual language expression in identifying financial fraud in both state-owned and non-state-owned firms.

The remainder of the article proceeds as follows. Section 2 reviews the theoretical framework and hypothesis development. Section 3 introduces the methods and variables selection, text indicators calculation, and sample sources and processing. Section 4 describes the empirical results and mechanisms. Section 5 provides robustness tests, and Section 6 concludes the paper.

2. Theoretical framework and hypothesis development

The financial information disclosure of listed firms serves as a bridge for public corporations to share information with investors and the general public in a comprehensive manner. Authentic, comprehensive, timely, and sufficient information disclosure is the basis for investing decisions. However, due to factors like motivation/pressure, opportunity, and justifications, the likelihood of firms submitting misleading information is increasing. Changes in disclosures can be extremely relevant to fraud because it is primarily motivated by deteriorating financial conditions and corporate performance (Rezaee, 2005). Misrepresentations can be used to conceal the misappropriation or misapplication of finances. Managers may submit incorrect financial reports to deceive investors or authorities about an enterprise's financial health and future prospects (Platt, 2015).

Studies have found that executive conference calls (Larcker and Zakolyukina, 2012), earnings press releases (Davis and Tama-Sweet, 2012), and annual reports (Loughran and McDonald, 2011) are important sources of text for mining to detect financial fraud. Annual reports are easy to obtain and can intuitively reflect the business situation, opportunities, and challenges faced by firms, making them the preferred choice for most researchers. Therefore, text information disclosed in financial reports has become critical complementary information for detecting financial fraud. With the gradual improvement in information disclosure norms, an increasing number of scholars have researched MD&A information disclosure. Zhang et al. (2022) use the bag-of-words (BOW) model to transform MD&A texts into digital data to detect financial fraud. In addition, using textual indicators as supplementary variables to identify financial fraud can improve the fraud recognition rate. Cecchini et al. (2010) increase the accuracy of fraud identification from 75.00% to 81.95% by combining textual and financial data. Based on the findings above, the following hypothesis is proposed.

Hypothesis 1. MD&A text indicators can effectively identify financial fraud.

We construct a text analysis framework using three aspects: textual language structure, language quality, and language expression. Meanwhile, language structure is divided into four parts: pronunciation, semantics, vocabulary, and grammar. Based on the genre of the collected text, we choose two aspects: semantics and vocabulary. The classifications of the text indicators are shown in Table 1.

Forward-looking vocabulary, particularly the extent of its information as used in MD&A texts, can measure a firm's prospects. The inclusion of forward-looking information in MD&A, such as investment plans, has a higher correlation with the future performance of the firm and a better predictive effect than historical information (Goodman et al., 2014). When stock price efficiency is low, the annual report includes more forward-looking MD&A texts to increase future information content of stock prices (Muslu et al., 2015). However, there is limited research on the application of forward-looking indicators to the identification of financial fraud. Some scholars believe that the adverbs associated with forward-looking information can indicate the quality and depth of information disclosure. The more adverbs are used, the higher the emotional intensity displayed, and the greater the likelihood of fraud in the "adverb modifies adjective" pattern (Goel and Uzuner, 2016).

Hypothesis 2. Higher redundancy of the language structure can expose financial fraud.

Risk indicators such as length, stickiness, redundancy, and specificity, can account for virtually all increases in fair value, internal controls, and risk factor disclosures (Dyer et al., 2017). MD&A texts by listed firms is used to compare the content and the information disclosed in the previous year with that in the current year, measured as the proportion of repeated text (words or phrases). The literature on the similarity of financial report texts in accounting can be traced back to Brown and Tucker (2011). They use the similarity of MD&A texts in annual financial reports as a measure of information content and report that capital markets respond positively to changes in MD&A texts. The greater the similarity, the greater the likelihood that the firm will cover up fraud. This is because in normal operations, firms adjust their operational content and plans from year to year, and excessive similarity reflects the management's unwillingness to disclose new information about the firm's operations, demonstrating the nature of the cover-up.

Based on the calculation concept of the similarity indicator, we propose a matching degree indicator. The matching degree of information density across different sections of the MD&A text disclosed by the listed firm can indicate the degree of detail in the information disclosure. Studies show that firms with a higher degree of matching across different texts provide more accurate information about the firm's core competitiveness (Muslu et al., 2015). To the best of our knowledge, the relationship between MD&A text-matching degree and corporate fraud identification has not been studied. However, based on accounting practices and the calculation concept of the similarity indicator, we believe that the lower the match between past operating results and future operating plans disclosed by a firm, the greater the likelihood of fraud cover-up. Based on the findings above, the following hypotheses are proposed.

Hypothesis 3a. A higher similarity of the language structure helps to identify financial fraud.

Hypothesis 3b. A higher matching degree of the language structure helps to identify financial fraud.

With the gradual improvement in information disclosure norms, an

Table 1
Text indicators classifications.

Classification	Fine classification	Text indicators
Language structure	Text vocabulary	Forward-looking
	Text semantics	Similarity
		Matching degree
Language quality		Readability
Language expression		Positive sentiment
		Negative sentiment

increasing number of scholars have researched the textual readability of MD&A information disclosures. Complex words, technical terms, and sentence length can affect readability. Understanding more complex accounting terminology requires additional knowledge from relevant personnel. Therefore, increasing the complexity of the text and enhancing the difficulty of extracting text information (thus hiding information content that the firm is unwilling to disclose) are common methods that firms use to manipulate earnings (Li, 2008). Lo et al. (2017) find that firms implementing accounting discretion and poor readability to conceal adverse information potential performance problems often exhibit more complex disclosures. Managers achieve the goal of earnings management by increasing the disclosure complexity and omitting negative information in MD&A texts (Paul and Sharma, 2023). Based on the findings above, the following hypothesis is proposed.

Hypothesis 4. Worse language quality foreshadows a higher possibility of financial fraud.

The tone expressed in the MD&A text by a listed firm may reflect its sentiments about past operations and future plans (Loughran and McDonald, 2011). Davis and Tama-Sweet (2012) use a program that automatically counts the frequency of positive and negative words to perform emotional statistical analyses of MD&A texts and find that MD&A tonal analysis can provide forward-looking information for future performance predictions. Murphy et al. (2018) examine industry recommendations and interviews with individuals experienced in writing the MD&A section and find that word choice and tone of the MD&A text can assist to detect fraudulent financial reports. On average, false MD&A text contains three times more positive and four times more negative emotions compared to true MD&A (Goel and Uzuner, 2016). Based on the findings above, the following hypotheses are proposed.

Hypothesis 5a. More positive intonation of the language expression helps to identify financial fraud.

Hypothesis 5b. More negative intonation of the language expression helps to identify financial fraud.

3. Research design

3.1. Empirical methods

Scholars have used different models to detect financial fraud. Lisic et al. (2015) use a logistic regression model based on financial ratios to detect financial fraud, while Rahimikia et al. (2017) use logistic regression, MLP NN, and a support vector machine to detect financial fraud. An explainable attention network can also be used in fraud detection (Farbmacher et al., 2022). Zhao et al. (2022) combine sentiment tone features extracted from comments in online stock forums, MD&A text, and financial statement notes, and use the CatBoost model to predict financial distress. Rahman and Zhu (2023) use AdaBoost, XGBoost, CUSBoost, and RUSBoost to detect financial fraud. Therefore, in this study, we choose three models—logistic regression, XGBoost algorithm, and an MLP NN—for the following reasons. Logistic regression is a classic model for financial fraud detection based on multiple dimensions (Rind et al., 2022); financial fraud still largely relies on traditional machine learning. XGBoost is suitable for algorithm and feature engineering. The use of neural networks has become good practice in the field of machine learning (we choose an MLP NN).

3.1.1. The logistic regression model

The logistic regression model is a classification model built on linear regression. We choose the binary logistic regression model to solve the binary classification problem, as follows:

$$p(y=1|X;\theta)=\frac{e^{\theta X}}{1+e^{\theta X}} \tag{1}$$

$$p(y=0|X;\theta)=\frac{1}{1+e^{\theta X}} \tag{2}$$

where X represents the independent variable, y represents the category to which the dependent variable belongs, and θ is the parameter to be determined by the model.

According to the models above and the predicted probability P, the logistic regression model is suitable for predicting financial fraud.

3.1.2. The XGBoost model

XGBoost is a supervised learning algorithm comprised of multiple decision trees that can solve machine-learning problems, such as regression and classification (Carmona et al., 2019). The XGBoost model is expressed as follows:

$$\hat{y}_i=\varphi(x_i)=\sum_{k=1}^{K} f_k(x_i), f_k\in F \tag{3}$$

For the given dataset, there are n samples and m features, defined as follows:

$$\mathbb{C}=\{(x_i,y_i)\}\ (|\mathbb{C}|=n, x_i\in\mathbb{R}^m, y_i\in\mathbb{R}) \tag{4}$$

$$F=\left\{f(x)=\omega_{q(x)}\right\}(q:\mathbb{R}^m\to \mathrm{T}, \omega\in\mathbb{R}^{\mathrm{T}}) \tag{5}$$

where $\hat{y}_i$ is the prediction result of the first sample during model training;, F represents the set of all decision trees in the model; f_k is one of the decision trees; q represents the score of the structure of each sample mapping tree to the corresponding leaf node; $\omega_{q(x)}$ represents the fractions of all leaf nodes of q in the set; and T is the number of leaf nodes of tree q. When inputting a sample, the sample is mapped to the leaf node according to the model and the predicted score.

According to the models above, the core idea of XGBoost is that the results of all weak classifiers add up to the predicted value, and the next weak classifier fits the gradient/residual of the error function of the predicted value (the error between the predicted value and the true value), thereby continuously reducing the residuals until the error requirements of the system are met. Thus, the XGBoost model is suitable for predicting financial fraud.

3.1.3. The MLP NN model

The MLP NN model is a feedforward artificial neural network model, which contains input, hidden, and output layers. The model maps multiple input datasets to a single output dataset, arbitrarily determines the initial value of the discriminant function, and gradually corrects data during the training process of sample classification until final confirmation (Rahimikia et al., 2017).

The input layer is represented by; the hidden layer is represented by h ; and the output layer is represented by o. For a small batch of samples $X\in R^{n\times d}$, the batch size is n and the number of inputs is d. Suppose that the MLP has only one hidden layer, whose output is H; then, $H\in R^{n\times d}$. Given that the hidden and output layers are fully connected, the weight and bias parameters of the hidden layer can be set as follows:

$$W_h\in R^{d\times h}, b_h\in R^{1\times h} \tag{6}$$

The weight and bias parameters of the output layer are as follows:

$$W_o\in R^{d\times h}, b_o\in R^{1\times h} \tag{7}$$

The output of the hidden layer is directly used as the input of the output layer. If Eq. (6) and Eq. (7) are combined, we obtain the following:

$$O=(XW_h+b_h)W_o+b_o=XW_hW_o+b_hW_o+b_o \tag{8}$$

According to the principles above, the MLP NN model will continuously learn through the training set to obtain more suitable weights

(parameters) and biases (thresholds) for the predicted value. Thus, the MLP NN model is suitable for predicting financial fraud.

3.2. Variables and indicators

3.2.1. Financial variables

Following relevant research, we select the following financial (ten) and corporate governance (three) variables: other receivables on total assets (*ORTA*), financial expense ratio (*FER*), gross margin index (*GMI*), tax to profit ratio (*TPR*), turnover of fixed assets (*TFA*), equity turnover (*ET*), growth ratio of total assets (*GRTA*), return on assets (*ROA*), operating net cash flow per share (*ONCS*), and operating net cash flow to financing expense (*ONCE*); *Dual*, *Top1*, and *Board*. Lennox et al. (2013) control the *ROA*, *Dual*, *Top1*, and *Board* variables. The *ET*, *GRTA* (*NRTA*; Bao et al., 2020), *GMI* (Dikmen and Küçükkocaoğlu, 2010), *TFA* (*FA*; Wang J. et al., 2021), and other financial variables are used in detecting financial fraud.

ORTA is the percentage of other receivables that comprise total assets. Other receivables include various amounts owed and temporary payments, which are highly inclusive, prone to fraud, and consequently can result in confusion and inconsistent reporting. Thus, further attention should be given to firms whose *ORTA* is high. For instance, *ORTA* was high for the analysis of the Jiangte Electric reports.[1]

FER is the percentage of financial expenses of main business income. Enterprises use *FER* to analyze the level of financial expenses incurred in financial operations and the flow of financial capital. When *FER* is significantly higher than the average level of comparable firms in the same industry, the financial risk is high. Thus, once the capital chain collapses, the firm faces a greater debt repayment risk. A very high *FER* was evident in the financials of Central South Construction.[2]

TPR is the percentage of business taxes and surcharges relative to the profit from business operation. Business taxes and surcharges, including consumption tax, resources tax, tax for maintaining and building cities, and education supplementary tax, are paid as value-added taxes. Empirical evidence from Chinese firms shows that *TPR* is lower if a firm is operating during a crisis of the firm's making.[3]

ONCS is the percentage of operating net cash flow over the share capital. *ONCS* reflects not only the support level of cash from the main business to the capital per share but also the ability of the listed firm to pay dividends. If the percentage of cash dividends distributed by the firm to the share capital exceeds *ONCS*, the firm needs to borrow money to pay dividends. Empirical evidence from Chinese firms shows that firms facing financial crisis are likely to commit financial fraud.[4]

ONCE is the percentage of operating net cash flow relative to debt interest and indicates a firm's ability to repay debt interest with cash obtained from operating activities. Empirical evidence from Chinese firms shows that if *ONCE* is low, the security and stability of debt repayment will decrease, which will affect the firm's normal production and operation and reduce profitability and future viability, which are all likely to be associated with financial fraud.[5]

3.2.2. Text indicators

Text mining was designed by DeJong (1982) in 1982. The earliest classification dictionary used in text mining research is the Harvard University General Survey (GI) Dictionary (Tetlock, 2007). However, the GI dictionary is not specific to financial text analysis. The Loughran–MacDonald (LM) dictionary, which uses positive and negative words, has been widely used in finance and accounting studies (Feldman et al., 2010). The early text mining classification dictionary developed in China (the *Cihai* Dictionary) is insufficient for financial and accounting research. Professional dictionaries, such as *Dalian University of Technology's Chinese Emotional Vocabulary Ontology Database* and *Chinese Sentiment Dictionary in the Financial Field*, have become more widely used in Chinese finance and accounting studies.

To digitize unstructured text, we use algorithms to calculate text indicators. First, we use Python crawler technology to automatically capture the annual reports of all listed firms from "www.cninfo.com" in China. The selected texts contain the final, revised, and complete annual reports of listed manufacturing firms in China. Reports in PDF form are then converted to TXT version. Next, the MD&A sections are manually extracted from the reports, and the MD&A texts are processed by sentence and word segmentation using the Python tool. Sentence segmentation is conducted according to punctuation, while word segmentation is carried out using *Jieba* word segmentation. This provides data for quantitative analysis of text indicators. The term frequency-inverse document frequency (TF-IDF) algorithm is also used in the word segmentation stage of the text. A high word frequency within a particular document and a low document frequency of that word in the set of documents can result in a highly weighted TF-IDF.

We choose six text indicators from language structure, language quality, and language expression to identify financial fraud.

Readability. The frequency of sentences containing accounting terminology in MD&A texts represents readability. The selected dictionary of accounting terminology is the *accounting terms glossary*, which constitutes *enterprise accounting standards and application guidelines*. We perform word segmentation based on *Jieba* and calculate the total number of words in the text, number of all professional accounting terms, and ratio.

Forward-looking. We use the frequency of degree adverbs of forward-looking information in MD&A texts to detect whether the text is forward-looking. We select the degree adverb table in the *Sentiment Analysis Dictionary*, developed from the *Dalian University of Technology Chinese Emotional Vocabulary Ontology Database.* This resource is constructed using Ekman's six categories of sentiment classification system. This dictionary is widely referenced in text analysis and is considered to have strong reliability (Diao et al., 2021). Similarly, we perform word segmentation and calculate the total number of words in the text, number of all degree adverbs, and ratio.

Similarity. Cosine similarity, a widely-known technique that works well for identifying similarities between textual documents, is used to calculate the similarity indicator (Li et al., 2020). The cosine angle value is used to compare the similarity of the MD&A text information disclosed in two consecutive years. We use *Jieba* to perform word segmentation, CountVectorizer to vectorize the class of text, and TfidfTransformer to preprocess the class. We then calculate the cosine distance of the MD&A texts in two consecutive years as the similarity indicator.

Matching degree. For the analysis, we focus on the matching degree of information presented in the "business review" and "future development prospects" sections of the MD&A texts for each firm. We perform word segmentation using *Jieba*, employ CountVectorizer to vectorize the class of text, and use TfidfTransformer to preprocess the class. We then calculate the cosine distance between the "business review" and "future development prospects" of the MD&A texts as the matching degree indicator.

Sentiment. The frequency of sentences containing positive and negative words in the MD&A texts represents the positive or negative sentiment value of the text. We use the *Chinese Sentiment Dictionary in the Financial Field,* which localizes the LM English dictionary for the Chinese context and solves the problem of poor applicability of generic sentiment dictionaries to financial scenarios. This dictionary is widely referenced in text analysis in finance literature (Du et al., 2022). We

[1] http://www.cninfo.com.cn/new/disclosure/detail?stockCode=002176&announcementId=1206099379&orgId=9900003697&announcementTime=2019-04-26.

[2] http://www.cninfo.com.cn/new/disclosure/detail?stockCode=000961&announcementId=1206161216&orgId=gssz0000961&announcementTime=2019-04-30.

[3] https://www.163.com/dy/article/G4OQQ1IJ0512B07B.html.

[4] https://zhidao.baidu.com/question/762895672134068204.html.

[5] https://wen.baidu.com/question/442569479109822924.html.

Table 2
Variable definitions.

Variable	Definitions
Financial fraud	0, 1 (*virtual variable*)
Text indicator (Explanatory variable)	Readability (*Rab*)
	Forward-looking (*Fwl*)
	Similarity (*Sim*)
	Matching degree (*Md*)
	Positive sentiment (*Pos*)
	Negative sentiment (*Neg*)
Financial indicator (Control variable)	Other receivables on total assets *(ORTA)*
	Financial expense ratio (*FER*)
	Gross margin index *(GMI)*
	Tax to profit ratio (*TPR*)
	Turnover of fixed assets (*TFA*)
	Equity turnover (*ET*)
	Growth ratio of total assets *(GRTA)*
	Return on assets *(ROA)*
	Operating net cash flow per share (*ONCS*)
	Operating net cash flow to financing expense (*ONCE*)
Corporate governance indicator (Control variable)	*Dual*
	Top1
	Board

* Italics represents variables.

perform word segmentation using *Jieba* and, using the sentiment dictionary, calculate the proportion of positive and negative sentiment words. The variable definitions are shown in Table 2.

3.3. Sample and data

3.3.1. Sample selection and processing

We use accounting data and MD&A texts in the 2020 accounting reports of listed manufacturing firms in the Chinese A-share market. The 2020 reporting period was the first year of the COVID-19 pandemic in China. According to the CSRC, the highest number of firms to date (n = 261) experienced operational difficulties during this period (REF).[6] Firms experiencing difficulties have a strong motivation to whitewash reports and falsify disclosures to maintain a resilient operational corporate image. After firms with missing values and special treatment are excluded from the sample, 183 remaining firms were identified as committing financial fraud during this period.

Reurink (2018) defined financial reporting fraud as either systematic manipulation or deliberate misstatement of financial report notes by management. This includes the deliberate omission of known contingent liabilities, debt contracts, and related party transactions. To detect financial fraud, we select 13 violations, including fraudulent listing, false records (misleading statements), insider trading, manipulation of stock price, illegal guarantees, and unauthorized changes in fund use, in combination with the 16 types of violations issued by the CSRC, Shenzhen Stock Exchange, Shanghai Stock Exchange, Ministry of Finance and other institutions. Most current studies use a matching technique to match non-fraudulent firms with the fraudulent firms based on year, size, and industry (Mayew et al., 2015; Craja et al., 2020). To ensure the robustness of the empirical results, drawing on Mayew's (2015) method, we use random matched samples of 396 non-financial fraudulent firms based on similar assets with that of the Python tool (30 firms achieve a 1:3 ratio and the others achieve a 1:2 ratio, resulting in n = 396). To ensure that the size of the matched firms is similar to that of the fraudulent firms in the sample, the total assets of randomly matched firms are not to be more than ±20% of the total assets of the fraudulent firm.

We adopt an out-of-sample prediction, randomly dividing the dataset by 7:3 using the random cross-validation method and select 0.3 as the test set to verify the accuracy of the model. Through five divisions, the random cross-validation method can greatly reduce the instability caused by random division. At the same time, through multiple divisions and multiple trainings, the corresponding evaluation results are obtained, and the average value is considered the final score, thus improving its generalization ability. Furthermore, given the heterogeneity of state-owned enterprises (SOE) and non-state-owned enterprises (non-SOE) in the literature (Zhong et al., 2017; Yao et al., 2020), we divide the sample into listed manufacturing SOEs and non-SOEs. The sample settings are shown in Table 3.

The corporate financial data in this study come from the China Stock Market and Accounting Research Database and the Wind Database. The MD&A texts are sourced from annual reports, which are mainly captured through "www.cninfo.com" using web crawler technology. Missing corporate data are supplemented by data available from the official websites of the listed firms.

3.3.2. Descriptive statistics

The statistical data used in this study include six text indicators, ten financial indicators, and three corporate governance indicators. After sorting and screening, the descriptive statistics for the final sample are shown in Table 4. The mean values and standard deviations of the indicators of fraudulent firms are consistent with those of non-fraudulent firms. When fraudulent firms adjust misreported information in the MD&A texts and the financial statement, the adjustment causes the dataset of fraudulent firms to converge to that of non-fraudulent firms, as the fraudulent reporting has been altered. The data show that the comparison of all indicators reflects the complexity and ability to conceal corporate financial fraud, highlighting how detecting fraud is difficult. Thus, we need additional models and methods to analyze the selected indicators.

The correlations among the text, financial, and corporate governance indicators are shown in Table 5. The table shows that the correlation of 19 indicators is not high, indicating solvency (*FER*, *ET*, *GRTA*, *ROA*, *ONCS* and *ONCE*), profitability (*GMI* and *TPR*), operating ability (*ORTA* and *TFA*), and the level of corporate governance (*Dual*, *Top1* and *Board*) of the firms.

4. Empirical results and analysis

4.1. The logistic regression model

The results from the financial fraud risk detection model of all manufacturing listed firms constructed by the logistics regression method are shown in Columns 1–3 of Table 6. The data show that the area under the curve (AUC) of the financial fraud risk detection model using only financial indicators is 74.37%, accuracy (ACC) is 73.33%, and precision (PRE) is 62.72%. Adding all text indicators to the model decreases the AUC by 0.20% but increases the ACC and PRE by 0.46% and 1.05%, respectively. Adding only readability, forward-looking, and positive sentiment instead of all the text indicators to the model increases the AUC, ACC, and PRE by 0.22%, 0.46%, and 1.05%., respectively. The data show that textual quality, forward-looking information, and positive sentiment improve the accuracy of financial fraud risk

Table 3
Sample selection.

Sample	All listed manufacturing fraudulent firms		All listed manufacturing non-fraudulent firms	
Type	SOE	Non-SOE	SOE	Non-SOE
N	23	160	137	259
Total N	183		396	

[6] The number of fraudulent firms disclosed in each year before 2020 is smaller, for example, the number in 2017 is 29, the number in 2018 is 24, the number in 2019 is 17, and the number in 2020 is 261. It can be seen that the difference in the number of fraud disclosures between years is too large, and the data are not comparable, so a longer sample interval is not considered.

Table 4
Descriptive statistics Notes: Coded as 1 for a fraudulent firm, 0 otherwise.

Variable	Mean		Standard deviation		max		min	
	1	0	1	0	1	0	1	0
Rab	0.303	0.295	0.028	0.027	0.367	0.373	0.237	0.228
Fwl	0.120	0.115	0.017	0.019	0.160	0.169	0.056	0.058
Sim	0.122	0.125	0.075	0.076	0.572	0.711	0.007	0.000
Md	0.408	0.424	0.132	0.122	0.888	0.894	0.158	0.160
Pos	0.141	0.147	0.028	0.030	0.223	0.255	0.065	0.060
Neg	0.019	0.019	0.008	0.008	0.045	0.046	0.002	0.000
ORTA	0.021	0.008	0.040	0.012	0.337	0.119	0.000	0.000
FER	0.036	0.012	0.109	0.022	1.238	0.121	−0.155	−0.085
GMI	0.379	0.676	16.781	35.526	93.956	442.093	−198.362	−549.041
TPR	0.173	0.146	0.738	0.581	8.698	7.019	−1.093	−4.678
TFA	3.346	16.113	3.133	225.726	18.749	4494.097	0.214	0.302
ET	0.574	0.510	0.743	0.640	5.944	8.178	0.020	0.008
GRTA	−2.697	−0.968	15.404	6.261	28.030	6.897	−162.155	−64.646
ROA	0.019	0.025	0.037	0.041	0.173	0.234	−0.134	−0.154
ONCS	0.546	0.674	0.876	0.967	4.562	7.624	−1.770	−1.560
ONCE	4.609	−12.612	135.133	161.322	1290.991	1348.962	−1148.410	−1378.435
Dual	0.519	0.235	0.501	0.424	1	1	0	0
Top1	0.296	0.336	0.131	0.140	0.812	0.749	0.084	0.065
Board	2.086	2.110	0.186	0.185	2.708	2.708	1.609	1.609

* Italics represents variables.

detection. Adding other text indicators to the model does not improve its accuracy. Thus, we conclude that it is sufficient to observe the three significant indicators to detect financial fraud in the sample.

The results from the financial fraud risk detection model of state-owned listed manufacturing firms constructed by the logistics regression method are shown in Columns 4–6 of Table 6. The results show that the AUC of the financial fraud risk detection model with only financial indicators is 73.51% and the ACC and PRE are 84.58% and 65.71%, respectively. Adding all text indicators to the model increases AUC by 0.12% and has no effect on the ACC or PRE. Adding only the readability, forward-looking, similarity, and positive sentiment indicators instead of all text indicators to the model increases the AUC by 0.10%, and has on effect on the ACC or PRE. The data show that textual quality, forward-looking information, similarity, and positive sentiment conveyed by the text are sufficient to improve the accuracy of financial fraud risk detection (although the ACC and PRE do not increase, while the AUC increases slightly). Adding other text indicators to the model does not improve its accuracy. Thus, we conclude that it is sufficient to observe the four significant indicators to detect financial fraud in state-owned listed manufacturing firms.

The results from the financial fraud risk detection model of non-state-owned listed manufacturing firms constructed by the logistics regression method are shown in Columns 7–9 of Table 6. The data show that the AUC of the financial fraud risk detection model with only financial indicators is 68.79%, the ACC and PRE are 67.14% and 62.16%, respectively. Adding all text indicators to the model decreases the AUC by 0.15% and increases the ACC and PRE by 0.32% and 0.77%, respectively. Adding only readability, forward-looking, matching degree, and positive sentiment indicators instead of all the text indicators decreases the AUC by 0.16% and increases the ACC and PRE by 1.22% and 0.48%, respectively. The data show that quality, forward-looking information, matching degree, and positive sentiment of the text improve the accuracy of financial fraud risk detection (although the AUC decreases slightly and the ACC and PRE increase by a larger margin). Adding other text indicators to the model does not improve its accuracy. Thus, we conclude that the four significant indicators of MD&A texts are sufficient for detecting financial fraud in non-state-owned listed manufacturing firms.

We then compare the nine columns of data in Table 6. First, after all text indicators or significant text indicators are added to the fraud risk detection models of the three samples, their accuracy improves. Second, the highest ratio is the ACC of state-owned listed manufacturing firms at 84.58%. Third, according to ownership theory, the essential differences between SOEs and non-SOEs lies in the nature of property rights. As the property rights of SOEs belong to the state, they have sufficient funds and market share but less competition pressure. Managers pay more attention to maintaining positive enterprise development and have less pressure to commit fraudulent behaviors when facing market pressure. As the property rights of non-SOEs belong to individuals, they are more competitive and more prone to market pressure. To ensure that firms are not delisted, their managers pay more attention to the realization of the previous year's expected performance in the current year and have a stronger incentive to defraud.

4.2. The XGBoost model

The results from the financial fraud risk detection model of all listed manufacturing firms constructed using the XGBoost model are shown in Columns 1–3 of Table 7. The data show that the AUC of the financial fraud risk detection model using only financial indicators is 74.13%, the ACC and PRE are 74.14% and 62.92%, respectively. Adding all the text indicators to the model increases the AUC, ACC, and PRE by 0.06%, 1.03%, and 3.41%, respectively. Adding only readability, forward-looking, and positive sentiment indicators instead of all the text indicators to the model improves the AUC, ACC, and PRE by 1.08%, 1.38%, and 2.92%, respectively. The data show that the quality, forward-looking information, and positive sentiment of the text significantly improve the accuracy of financial fraud risk detection. Adding other text indicators to the model does not improve its accuracy. Thus, to detect the financial fraud of all listed manufacturing firms, it is sufficient to observe the three significant indicators.

The results from the financial fraud risk detection model of listed manufacturing SOEs constructed using XGBoost are shown in Columns 4–6 of Table 7. The data show that the AUC of the financial fraud risk detection model using only financial indicators is 79.51% the ACC and PRE are 86.25% and 64.76%, respectively. Adding all the text indicators to the model increases the AUC and PRE by 5.20% 16.91%, respectively, and has no effect on the ACC. Adding only readability, similarity, and positive sentiment indicators instead of all text indicators increases the AUC and PRE by 5.14% and 5.24%, respectively, leaving the ACC unaffected. The data show that quality, similarity, and positive sentiment of the text significantly improve the accuracy of financial fraud risk detection. Adding other text indicators to the model does not improve its accuracy. Thus, the three significant indicators are sufficient to detect the financial fraud of listed manufacturing SOEs.

The results from the financial fraud risk detection model of listed

Table 5
Correlation matrix.

variable	*Rab*	*Fwl*	*Sim*	*Md*	*Pos*	*Neg*	*ORTA*	*FER*	*GMI*	*TPR*	*TFA*	*ET*	*GRTA*	*ROA*	*ONCS*	*ONCE*	*Dual*	*Top1*	*Board*
Rab	1	−0.341	0.127	0.050	0.111	−0.227	0.074	0.044	0.000	−0.031	−0.065	0.027	0.003	0.038	0.068	−0.010	0.010	0.086	0.035
Fwl	−0.341	1	0.078	0.002	0.334	0.155	−0.004	−0.067	0.013	0.022	−0.021	0.017	0.013	−0.085	−0.035	0.017	−0.097	−0.044	0.118
Sim	0.127	0.078	1	−0.254	−0.001	0.088	−0.064	0.003	0.006	0.001	−0.002	−0.002	−0.020	0.101	0.001	0.011	0.000	0.013	−0.087
Md	0.050	0.002	−0.254	1	0.059	−0.223	−0.034	−0.008	−0.046	0.083	0.096	−0.082	−0.009	−0.012	0.033	−0.004	0.103	−0.009	0.053
Pos	0.111	0.334	−0.001	0.056	1	0.218	−0.059	−0.017	−0.041	−0.028	0.024	0.000	−0.031	0.010	−0.072	−0.036	0.056	−0.008	−0.014
Neg	−0.227	0.155	0.088	−0.223	0.218	1	−0.003	−0.044	0.044	−0.020	0.094	−0.076	0.077	0.093	−0.015	−0.054	0.070	−0.093	−0.096
ORTA	0.074	−0.004	−0.064	−0.034	−0.059	−0.003	1	−0.171	−0.039	−0.025	−0.103	−0.086	−0.023	0.060	0.075	0.021	0.104	0.062	−0.043
FER	0.044	−0.067	0.003	−0.008	−0.017	−0.044	−0.171	1	−0.017	−0.026	0.180	−0.012	0.059	0.121	0.028	−0.039	−0.083	0.115	0.008
GMI	0.000	0.013	0.006	−0.046	−0.041	0.044	−0.039	−0.017	1	−0.015	−0.003	−0.006	−0.020	−0.006	0.011	0.003	−0.043	−0.079	0.033
TPR	−0.031	0.022	0.001	0.083	−0.028	−0.020	−0.025	−0.026	−0.015	1	0.027	0.018	−0.030	0.049	0.039	−0.010	0.085	−0.023	0.006
TFA	−0.065	−0.021	−0.002	0.096	0.024	0.094	−0.103	0.180	−0.003	0.027	1	−0.241	0.006	0.005	0.034	−0.011	0.009	0.016	0.041
ET	0.027	0.017	−0.002	−0.082	0.000	−0.076	−0.086	−0.012	−0.006	0.018	−0.241	1	0.040	−0.204	−0.115	−0.069	−0.156	0.008	−0.118
GRTA	0.003	0.013	−0.020	−0.009	−0.031	0.077	−0.023	0.059	−0.020	−0.030	0.006	0.040	1	−0.229	−0.022	−0.016	0.051	0.042	−0.017
ROA	0.038	−0.085	0.101	−0.012	0.010	0.093	0.060	0.121	−0.006	0.049	0.005	−0.204	−0.229	1	−0.269	0.074	−0.053	−0.161	−0.056
ONCS	0.068	−0.035	0.001	0.033	−0.072	−0.015	0.075	0.028	0.011	0.039	0.034	−0.115	−0.022	−0.269	1	−0.100	0.047	−0.030	0.014
ONCE	−0.010	0.017	0.011	−0.004	−0.036	−0.054	0.021	−0.039	0.003	−0.010	−0.011	−0.069	−0.016	0.074	−0.100	1	0.050	0.033	0.088
Dual	0.010	−0.097	0.000	0.103	0.056	0.070	0.104	−0.083	−0.043	0.085	0.009	−0.156	0.051	−0.053	0.047	0.050	1	0.079	0.085
Top1	0.086	−0.044	0.013	−0.009	−0.008	−0.093	0.062	0.115	−0.079	−0.023	0.016	0.008	0.042	−0.161	−0.030	0.033	0.079	1	0.085
Board	0.035	0.118	−0.087	0.053	−0.014	−0.096	−0.043	0.008	0.033	0.006	0.041	−0.118	−0.017	−0.056	0.014	0.088	0.085	0.085	1

Notes: These variables are defined in Table 1 for all listed manufacturing firms used in this study.

manufacturing non-SOEs constructed using XGBoost are shown in Columns 7–9 of Table 7. The data show that using only financial indicators the AUC, ACC, and PRE of the financial fraud risk detection model are 71.35%, 68.41%, and 62.10%, respectively. Adding all the text indicators into the model decreases the AUC by 4.24% but increases the ACC and PRE by 0.48% and 2.59%, respectively. Adding only readability, forward-looking, matching degree, and positive sentiment indicators instead of all the text indicators decreases the AUC by 0.68%, but increases the ACC and PRE by 1.91% and 3.95%, respectively. The data show that quality, forward-looking information, matching degree, and positive sentiment of the text significantly improve the accuracy of financial fraud risk detection (although the AUC decreases, and both the ACC and PRE increase). Adding other text indicators to the model does not improve its accuracy. Thus, to detect financial fraud, it is sufficient to observe the four significant indicators of non-state-owned listed manufacturing firms.

We then compare the data in the nine columns in Table 7. First, after all text indicators or the significant text indicators are added to the fraud risk detection models of the three samples, the AUC, ACC, and PRE are significantly improved. Second, the AUC and ACC for fraud detection in state-owned listed manufacturing firms are higher than in the other samples at 84.71% and 86.25%, respectively. Third, according to ownership theory and principal–agent theory, the significant text indicators differ between the two types of firms. The reasons for these differences are explained in Section 4.1.

4.3. The MLP NN model

The results from the financial fraud risk detection model of all listed manufacturing firms constructed by the MLP NN model are shown in Columns 1–3 of Table 8. Using only the financial indicators, the AUC, ACC, and PRE of the financial fraud risk detection model are 57.42%, 65.17%, and 51.43%, respectively. Adding all text indicators to the model increases the AUC, ACC, and PRE by 6.64%, 4.83%, and 29.56%, respectively. When only the forward-looking, matching degree, and positive sentiment indicators are added to the model, the AUC increases by 2.26%, but the ACC and PRE increases by 4.03% and 5.83%, respectively. The data reveal that the forward-looking information, matching degree, and positive sentiment of the text significantly improve the accuracy of financial fraud risk detection. Adding other text indicators to the model does not improve its accuracy. Thus, the three significant indicators are sufficient to detect financial fraud in all listed manufacturing firms.

The results from the financial fraud risk detection model of listed manufacturing SOEs constructed using the MLP NN model are shown in Columns 4–6 of Table 8. The data indicate that the AUC, ACC, and PRE of the financial fraud risk detection model using only financial indicators are 62.80%, 80.42%, and 26.67%, respectively. Adding all text indicators into the model increases the AUC and ACC by 7.83% and 1.25%, respectively, but decreases the PRE by 16.67%. Adding only the forward-looking, matching degree, and positive sentiment indicators increases the AUC, ACC, and PRE by 29.43%, 7.50%, and 59.33%, respectively. The data indicate that the forward-looking information, matching degree, and positive sentiment of the text significantly improve the accuracy of financial fraud risk detection (although the PRE in Group 1 and Group 2 are low, that in Group 3 is high). Adding other text indicators to the model does not improve its accuracy. Thus, observing the three significant indicators is sufficient to detect financial fraud in state-owned listed manufacturing firms.

The results from the financial fraud risk detection model of listed manufacturing non-SOEs constructed by the MLP NN model are shown in Columns 7–9 of Table 8. The AUC, ACC, and PRE of the financial fraud risk detection model using only financial indicator variables are 62.59%, 66.03%, and 57.35%, respectively. Adding all text indicators to the model increases the AUC, ACC, and PRE by 1.55%, 3.49%, and 5.71%, respectively. Adding only readability, forward-looking, matching

Table 6
Results of the Logistic regression model.

Model	All			SOE			Non-SOE		
	AUC (1)	ACC (2)	PRE (3)	AUC (4)	ACC (5)	PRE (6)	AUC (7)	ACC (8)	PRE (9)
Group1	0.7437	0.7333	0.6272	0.7351	0.8458	0.6571	0.6879	0.6714	0.6216
Group2	0.7417	0.7379	0.6377	0.7363	0.8458	0.6571	0.6864	0.6746	0.6293
Group3(+ *Rab* + *Fwl* + *Pos*)	0.7459	0.7379	0.6377						
Group4(+ *Rab* + *Fwl* + *Sim* + *Pos*)				0.7361	0.8458	0.6571			
Group5(+ *Rab* + *Fwl* + *Md* + *Pos*)							0.6863	0.6762	0.6338

Notes: Group1 = *ORTA* + *FER* + *GMI* + *TPR* + *TFA* + *ET* + *GRTA* + *ROA* + *ONCS* + *ONCE* + *Dual* + *Top1* + *Board* = *Fin*Group2 = *Rab* + *Fwl* + *Sim* + *Md* + *Pos* + *Neg* + *Fin*Group3 = Text significant indicators for each sample + *Fin*These variables are defined in Table 2.AUC—area under the ROC curve; ACC—accuracy; PRE—prediction.

Table 7
Results of the XGBoost model.

Model	All			SOE			Non-SOE		
	AUC (1)	ACC (2)	PRE (3)	AUC (4)	ACC (5)	PRE (6)	AUC (7)	ACC (8)	PRE (9)
Group1	0.7413	0.7414	0.6292	0.7951	0.8625	0.6476	0.7135	0.6841	0.6210
Group2	0.7419	0.7517	0.6633	0.8471	0.8625	0.8167	0.6711	0.6889	0.6469
Group3(+ *Rab* + *Fwl* + *Pos*)	0.7521	0.7552	0.6584						
Group4(+ *Rab* + *Sim* + *Pos*)				0.8465	0.8625	0.7000			
Group5(+ *Rab* + *Fwl* + *Md* + *Pos*)							0.7067	0.7032	0.6605

Notes: See Table 6.

Table 8
Results of the MLP NN model.

Model	All			SOE			Non-SOE		
	AUC (1)	ACC (2)	PRE (3)	AUC (4)	ACC (5)	PRE (6)	AUC (7)	ACC (8)	PRE (9)
Group1	0.5742	0.6517	0.5143	0.6280	0.8042	0.2667	0.6259	0.6603	0.5735
Group2	0.6406	0.7000	0.8099	0.7063	0.8167	0.1000	0.6414	0.6952	0.6306
Group3(+ *Fwl* + *Md* + *Pos*)	0.5968	0.6920	0.5726						
Group4(+ *Fwl* + *Md* + *Pos*)				0.9223	0.8792	0.8600			
Group5(+ *Rab* + *Fwl* + *Md* + *Pos*)							0.6352	0.6810	0.5857

Note(s): See Table 6.

degree, and positive sentiment increases the AUC, ACC, and PRE by 0.93%, 2.07%, and 1.22%, respectively. The data reveal that the quality, forward-looking information, matching degree, and positive sentiment of the text improve the accuracy of financial fraud risk detection. Adding other text indicators to the model does not improve its accuracy. Thus, the four significant indicators are adequate for detecting financial fraud in listed manufacturing non-SOEs.

We compare the data across the columns in Table 8. First, after the significant text indicators are added to the fraud risk detection models of the three samples, the AUC, ACC, and PRE are significantly improved. Second, the AUC of fraud detection of listed manufacturing SOEs is higher (92.23%) than that of the other samples. For an explanation, see the analy sis in Section 4.1. Third, according to ownership theory and principal–agent theory, compared with state-owned listed manufacturing SOEs, non-SOEs try to make the report content obscure when preparing accounting reports to cover up the fact that they are inexperienced.

MD&A text indicators can effectively identify financial fraud. Thus, Hypothesis 1 is validated.

4.4. Comparison of the three models

We compare the three methods and present the results in Table 9. First, the readability and positive sentiment indicators are the common significant text indicators identified using logistic regression and the XGBoost model, whereas the forward-looking and positive sentiment indicators are those that are identified using logistic regression and the MLP NN model. For all the samples, the common significant text indicators of the three models are forward-looking and positive sentiment. For the SOEs sample, the most common significant text indicator is positive sentiment. For the non-SOEs sample, the common significant text indicators are readability, forward-looking, matching degree, and positive sentiment. Overall, the XGBoost model has the highest and most stable recognition accuracy, followed by the MLP NN model, with the logistic regression model exhibiting the lowest recognition accuracy.

Table 9 also presents the validation results. We select at least two repetitive textual indicators from the three models, which are

Table 9
Comparison of logistic, XGBoost, and MLP NN models.

Model	Recognition accuracy ranking	Enterprise type	Significant indicators	Unified indicators
Logistic	3	All	*Rab, Fwl, Pos*	*Rab, Fwl, Pos*
		SOE	*Rab, Fwl, Sim, Pos*	
		Non-SOE	*Rab, Fwl, Md, Pos*	
XGBoost	1	All	*Rab, Fwl, Pos*	*Rab, Pos*
		SOE	*Rab, Sim, Pos*	
		Non-SOE	*Rab, Fwl, Md, Pos*	
MLP NN	2	All	*Fwl, Md, Pos*	*Fwl, Md, Pos*
		SOE	*Fwl, Md, Pos*	
		Non-SOE	*Rab, Fwl, Md, Pos*	

Table 10
Regression for the impacts of text indicators on financial fraud.

Variables	Fraud
Rab	0.740***
	(0.244)
Fwl	0.518***
	(0.167)
Sim	−0.324
	(0.678)
Md	−1.576
	(1.119)
Pos	−0.618**
	(0.263)
Neg	0.120
	(0.071)
_cons	−2.814*
	(1.517)
N	579
R^2	0.036

Notes: Standard errors are in parentheses. Asterisks denote significance at the * = 0.10, ** = 0.05, and *** = 0.01 levels.

characterized by readability, forward-looking, and positive sentiment. Hypothesis 2, 4, and 5a have been preliminarily validated, but the mechanism has not yet been validated. Hypothesis 3a, 3b, and 5b have not been validated.

4.5. Mechanism

We further analyze the specific impact of the significant text indicators reported in Section 4.4 on financial fraud and explore their mechanisms. As the three models screened slightly differently for significant indicators, we analyze the text indicators (readability, forward-looking, and positive sentiment) that have a significant impact on financial fraud identification.

Table 10 reveals that the readability, forward-looking, and positive sentiment indicators are statistically significant. Furthermore, with the positive and negative coefficients, we find that the readability and forward-looking indicators are positively related to financial fraud, whereas the positive sentiment is negatively related to financial fraud.

As such, complex vocabulary, technical terms, and sentence length can affect readability. Increasing the complexity of the text and enhancing the difficulty of extracting text information have become common methods for masking financial fraud. According to information asymmetry theory, the degree of information mastery between entities varies, as they are always willing to take risks to maximize their own interests (George, 1970). If the accounting terminology used in a firm's annual report is too professional and complex, making it difficult for report users to understand, we can assume that this is intended to conceal financial fraud. Moreover, a high use of degree adverbs about forward-looking information in the MD&A text is associated with firms motivated to commit financial fraud. Fraudulent firms might modify the wording to express more exaggerated degrees that do not have high disclosure value. This leads to the principal–agent problem. As two common issues in principal–agent theory, adverse selection and moral hazard can utilize information to participate in self-profit conduct (Jensen and Meckling, 1976). Exaggerated disclosure of forward-looking information causes adverse selection problems and increases the risk of financial fraud. The literature reveals that more positive sentiments are associated with a higher likelihood of financial fraud. However, we draw a contrary conclusion: the less positive the intonation of the language expression, the higher the likelihood of financial fraud. As firms experienced a severe systemic external crisis in 2020, the first year of the COVID-19 pandemic in China, they were poorly run and resorted to fraudulent behaviors. According to stakeholder theory (Freeman, 1984), firms tend to pursue profits at the expense of their employees, customers, creditors, and even the government when faced with an external crisis. Coupled with the more cautious corporate culture characteristic of China, firms do not show much positive emotion in their financial reports.

The significant financial fraud identification indicators screened by the three models are consistent with the financial fraud identification indicators evaluated using practical experience. Therefore, stakeholders should pay attention to the professionalism of accounting terms and the complexity of the wording, observe the degree of tone emphasis on forward-looking information, and determine whether positive emotions are restrained in the MD&A text of financial reports.

Thus, we further verify that higher redundancy of language structure and worse language quality help to identify financial fraud as shown in Table 10. Thus, Hypotheses 2 and 4 are validated. However, less positive intonation of the language expression can detect financial fraud. In the case of Hypothesis 5a, we come to the opposite conclusion.

5. Robustness tests

In this section, we present two supplemental tests for checking the robustness of the results. We use a new method to measure the readability and sentiment indicators and replace the original indicators for fraud identification. Forward-looking indicators do not have alternative variables.

According to Hypothesis 4, worse language quality foreshadows a higher possibility of financial fraud. We use the ratio of common words, professional words, conjunction words, and average sentence length in the MD&A texts to recalculate the readability, which is defined as the follows:

$$Rab2 = 1 - common\ words\ /\ total\ words + professional\ words\ /\ total\ words + conjunction\ words\ /\ total\ words + average\ sentence\ length\ /\ total\ sentences \tag{9}$$

Then, we normalize the four ratios as follows:

$$mmx\_common = [(1 - common\ words) - min(1 - common\ words)]/[max(1 - common\ words) - min(1 - common\ words)] \tag{10}$$

$$mmx.professional = [(1 - professional\ words) - min(1 - professional\ words)]/[max(1 - professional\ words) - min(1 - professional\ words)] \tag{11}$$

$$mmx.conjunction = [(1 - conjunction\ words) - min(1 - conjunction\ words)]/[max(1 - conjunction\ words) - min(1 - conjunction\ words)] \tag{12}$$

$$mmx\_average\ sentence = [(1 - average\ sentence\ length) - min(1 - average\ sentence\ length)]/[max(1 - average\ sentence\ length) - min(1 - average\ sentence\ length)] \tag{13}$$

The larger the indicator, the worse the readability and the higher the possibility of fraud. We check the robustness of readability in the MD&A texts using the *Rab2* variable (Column 1, Table 11). Consistent with the results above, the *Rab2* indicator is significant, revealing that the positive effect of *Rab2* on financial fraud is robust to the alternative measure of *Rab*.

According to Hypothesis 5a and the verification result, less positive intonation of language expression helps to identify financial fraud. We use the positive and negative words in the MD&A texts to recalculate intonation, which is defined as follows:

$$IMT = (the\ number\ of\ positive\ words - the\ number\ of\ negative\ words)/total\ words\ of\ MD\&A \tag{14}$$

The larger the indicator, the more positive the intonation. We check

Table 11
Robustness test for text indicators on financial fraud.

Variables	Fraud	Fraud
	(1)	(2)
Rab		0.740***
		(0.244)
Rab2	1.727**	
	(0.835)	
Fwl	0.518***	0.518***
	(0.167)	(0.167)
Sim	−0.324	−0.324
	(0.678)	(0.678)
Md	−1.576	−1.576
	(1.119)	(1.119)
Pos	−0.618**	
	(0.263)	
Neg	0.120	
	(0.071)	
IMT		−0.423*
		(0.180)
_cons	−1.801	−3.062**
	(1.389)	(1.197)
N	579	579
R^2	0.032	0.040

Notes: Standard errors are in parentheses. Asterisks denote significance at the * = 0.10, ** = 0.05, and *** = 0.01 levels.
* Italics represents variables.

the robustness of the intonation in the MD&A texts using the *IMT* variable (Column 2, Table 11). Consistent with the results above, the *IMT* indicator is significant, revealing that the negative effect of *IMT* on financial fraud is robust to the alternative measures of the positive sentiment indicator. Overall, these additional results are consistent with the main findings.

6. Conclusions

Using a sample of listed manufacturing firms in China, we analyze the MD&A texts of 183 fraudulent firms and 396 non-fraudulent firms. Six text disclosure indicators are selected from three language dimensions: structure, quality, and expression, to form a framework. We detect the significant textual indicators in MD&A texts for this sample and compare the detection capabilities of the financial fraud models before and after adding text indicators.

The results demonstrate that text indicators can increase the accuracy of detecting financial fraud compared to using only financial indicators. After six MD&A text indicators are added to the financial fraud detection model, the financial fraud recognition rate improves significantly. Among the framework indicators, readability, forward-looking, and positive sentiment are significant, indicating that the structure, quality, and expression of the language of the MD&A texts can identify financial fraud. In addition, readability and forward-looking are positively correlated with financial fraud, whereas positive sentiment is negatively correlated with financial fraud. This indicates that poorer language quality, higher redundancy of the language structure, and less positive intonation of the language expression identifies financial fraud. Furthermore, after text indicators are added to financial indicators, the recognition accuracy of the models is generally improved. The XGBoost model has the best recognition accuracy among the three models, and the logistic model has the lowest recognition accuracy. Finally, positive sentiment is the common significant text indicator across the three models for the SOE sample, whereas readability, forward-looking, matching degree, and positive sentiment are the common significant indicators for the non-SOE sample. According to ownership theory, managers hold an optimistic attitude towards the SOE's operation, and fraudulent behavior is simple and easy to identify. Managers of non-SOEs have more experience in fraudulent behavior to prevent their firms from being delisted, making it more difficult to identify fraud. Thus, the positive sentiment indicator highlights the importance of considering textual language expression in detecting financial fraud in SOEs and non-SOEs.

Our findings provide several useful implications. First, textual information can complement the efficiency of financial indicators for financial fraud identification of listed firms, expand the dimensions and sources of textual information, and screen indicators with significant impact from them, improve the accuracy of financial fraud identification, reduce information asymmetry, and expand the stakeholder theory. Second, an empirical analysis is conducted with three models and robust data from the Chinese A-share market to test the feasibility of textual indicators to identify financial fraud and provide references for stakeholders' practice and decision making.

Owing to the limitations of text information, such as the lack of ex-post verification, listed firms have more options to disclose text information. Thus, a larger set of MD&A texts is required and more dimensions and sources of text indicators should be considered to maximize the value of text information and improve the financial fraud detection of listed firms. Furthermore, text mining algorithms for MD&A texts in financial reports can be optimized to improve the performance of financial fraud prediction models.

The authors are grateful to the editor, the guest editor, and two anonymous referees for their thoughtful comments and helpful guidance. They also thank participants at the Digital Innovation and Financial Access for Enterprises conference held at Shandong Normal University. The usual disclaimer applies.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The authors acknowledge financial support from National Natural Science Foundation of China (72003110).

References

Bao, Y., Ke, B., Li, B., Yu, Y.J., Zhang, J., 2020. Detecting accounting fraud in publicly traded US firms using a machine learning approach. J. Account. Res. 58 (1), 199–235.
Brown, S.V., Tucker, J.W., 2011. Large-sample evidence on firms' year-over-year MD&A modifications. J. Account. Res. 49 (2), 309–346.
Carmona, P., Climent, F., Momparler, A., 2019. Predicting failure in the U.S. banking sector: an extreme gradient boosting approach. Int. Rev. Econ. Finance 61, 304–323.
Cecchini, M., Aytug, H., Koehler, G.J., Pathak, P., 2010. Detecting management fraud in public companies. Manag. Sci. 56 (7), 1446–1160.
Craja, P., Kim, A., Lessmann, S., 2020. Deep learning for detecting financial statement fraud. Decis. Support Syst. 139, 113421.
Davis, A.K., Tama-Sweet, I., 2012. Managers' use of language across alternative disclosure outlets: earnings press releases versus MD&A. Contemp. Account. Res. 29 (3), 804–837.
DeJong, G.F., 1982. An overview of the frump system. In: Lehnert, W.G., Ringle, M.H. (Eds.), Strategies for Natural Language Processing. Erlbaum, Hillsdale, N.J.
Diao, Y.F., Lin, H.F., Yang, L., Fan, X.C., Chu, Y.H., Wu, D., Xu, K., 2021. Emotion cause detection with enhanced representation attention convolutional context network. Soft Comput. 25, 1297–1307.
Dikmen, B., Küçükkocaoğlu, G., 2010. The detection of earnings manipulation: the three-phase cutting plane algorithm using mathematical programming. J. Forecast. 29 (5), 442–466.
Durnev, A., Mangen, C., 2020. The spillover effects of MD&A disclosures for real investment: the role of industry competition. J. Account. Econ. 70 (1), 101299.
Du, Z.J., Huang, A.G., Wermers, R., Wu, W.F., 2022. Language and domain specificity: a Chinese financial sentiment dictionary. Rev. Finance 26 (3), 673–719.
Dyck, A., Morse, A., Zingales, L., 2010. Who blows the whistle on corporate fraud? J. Finance 65 (6), 2213–2253.

Dyer, T., Lang, M., Stice-Lawrence, L., 2017. The evolution of 10-K textual disclosure: evidence from latent dirichlet allocation. J. Account. Econ. 64 (2–3), 221–245.
Farbmacher, H., Low, L., Spindler, M., 2022. An explainable attention network for fraud detection in claims management. J. Econom. 228 (2), 244–258.
Feldman, R., Govindaraj, S., Livnat, J., Segal, B., 2010. Management's tone change, post earnings announcement drift and accruals. Rev. Account. Stud. 15, 915–953.
Freeman, R.E., 1984. Strategic Management: A Stakeholder Approach. Pitman Press, Boston.
George, A.A., 1970. The market for 'lemons': quality uncertainty and the market mechanism. Q. J. Econ. 84 (3), 488–500.
Goel, S., Uzuner, O., 2016. Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports. Intell. Syst. Account. Finance Manag. 23 (3), 215–239.
Goodman, T.H., Neamtiu, M., Shroff, N., White, H.D., 2014. Management forecast quality and capital investment decisions. Account. Rev. 89 (1), 331–365.
Gillan, S.L., Martin, J.D., 2007. Corporate governance post-Enron: effective reforms, or closing the stable door? J. Corp. Finance 13 (5), 929–958.
Jensen, M.C., Meckling, W.H., 1976. Theory of the firm: managerial behavior, agency costs, and ownership structure. J. Financ. Econ. 3 (4), 305–360.
Kong, D., Shi, L., Zhang, F., 2021. Explain or conceal? Causal language intensity in annual report and stock price crash risk. Econ. Modell. 94, 715–725.
Larcker, D., Zakolyukina, A.A., 2012. Detecting deceptive discussions in conference calls. J. Account. Res. 50 (2), 495–540.
Lennox, C.S., Lisowsky, P., Pittman, J.A., 2013. Tax aggressiveness and accounting fraud. J. Account. Res. 51 (4), 739–778.
Li, F., 2008. Annual report readability, current earnings, and earnings persistence. J. Account. Econ. 45 (2–3), 221–247.
Li, J.Y., Li, J.P., Zhu, X.Q., 2020. Risk dependence between energy corporations: a text-based measurement approach. Int. Rev. Econ. Finance 68, 33–46.
Lisic, L.L., Sliveri, S., Song, Y., Wang, K., 2015. Accounting fraud, auditing, and the role of government sanctions in China. J. Bus. Res. 68 (6), 1186–1195.
Lo, K., Ramos, F., Rogo, R., 2017. Earnings management and annual report readability. J. Account. Econ. 63, 1–25.
Loughran, T., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. J. Finance 66 (1), 35–65.
Maharjan, J., Lee, S.W., 2021. Short-selling pressure and year-over-year MD&A modifications. Account. Finance 62 (3), 3513–3562.
Mayew, W.J., Sethuraman, M., Venkatachalam, M., 2015. MD&A disclosure and the firm's ability to continue as a going concern. Account. Rev. 90 (4), 1621–1651.
Murphy, P.R., Purda, L., Skillicorn, D., 2018. Can fraudulent cues be transmitted by innocent participants? J. Behav. Finance 19 (1), 1–15.
Muslu, V., Radhakrishnan, S., Subramanyam, K.R., Lim, D., 2015. Forward-looking MD&A disclosures and the information environment. Manag. Sci. 61 (5), 931–948.
Paul, S., Sharma, P., 2023. Does earnings management affect linguistic features of MD&A disclosures? Finance Res. Lett. 51, 103352.
Platt, S., 2015. Criminal Capital: How the Finance Industry Facilitates Crime. Palgrace Macmillan, Basingstoke.
Que, J.J., Zhang, X.Y., 2019. Pre-IPO growth, venture capital, and the long-run performance of IPOs. Econ. Modell. 81, 205–216.
Rahimikia, E., Mohammadi, S., Rahmani, T., Ghazanfari, M., 2017. Detecting corporate tax evasion using a hybrid intelligent system: a case study of Iran. Int. J. Account. Inf. Syst. 25, 1–17.
Rahman, M.J., Zhu, H.T., 2023. Predicting accounting fraud using imbalanced ensemble learning classifiers - evidence from China. Account. Finance, 13044.
Rezaee, Z., 2005. Causes, consequences, and deterrence of financial statement fraud. Crit. Perspect. Account. 16 (3), 277–298.
Reurink, A., 2018. Financial Fraud: a literature review. J. Econ. Surv. 32 (5), 1292–1325.
Rind, A.A., Abbassi, W., Allaya, M., Hammouda, A., 2022. Local peers and firm misconduct: the role of sustainability and competition. Econ. Modell. 116, 106000.
Tetlock, P.C., 2007. More than words: quantifying language to measure firms' "fundamentals". J. Finance 63 (3), 1437–1467.
Wang, J., Li, J., Zhang, Q.J., 2021. Does carbon efficiency improve financial performance? Evidence from Chinese firms. Energy Econ., 105658
Wang, L., Chen, X., Li, X., Tian, G., 2021. MD&A readability, auditor characteristics, and audit fees. Account. Finance 61, 5025–5050.
Yao, S.Y., Wang, Z.Q., Sun, M.Y., Liao, J., Cheng, F.Y., 2020. Top executives' early-life experience and financial disclosure quality: impact from the great Chinese famine. Account. Finance 60 (5), 4757–4793.
Zhang, Y., Hu, A.L., Wang, J.H., Zhang, Y.J., 2022. Detection of fraud statement based on word vector: evidence from financial companies in China. Finance Res. Lett. 46, 102477.
Zhao, S., Xu, K., Wang, Z., Liang, C., Lu, W., Chen, B., 2022. Financial distress prediction by combining sentiment tone features. Econ. Modell. 106, 105709.
Zhong, Q.L., Liu, Y.Y., Yuan, C., 2017. Director interlocks and spillover effects of board monitoring: evidence from regulatory sanctions. Account. Finance 57 (5), 1605–1633.
Zhu, X., Ao, X., Qin, Z., Chang, Y., Liu, Y., He, Q., Li, J., 2021. Intelligent financial fraud detection practices in post-pandemic era. Innovation (Cambridge (Mass)) 2 (4), 100176.