Received: 21 January 2022 | Accepted: 26 August 2022

DOI: 10.1111/poms.13888

POMS

ORIGINAL ARTICLE

Tracking disclosure change trajectories for financial fraud detection

Rong Liu[1] | **Jujun Huang**[1] | **Zhongju Zhang**[2]

[1]School of Business, Stevens Institute of Technology, Hoboken, New Jersey, USA

[2]W.P. Carey School of Business, Arizona State University, Tempe, Arizona, USA

Correspondence
Rong Liu, School of Business, Stevens Institute of Technology, Hoboken, NJ, USA.
Email: rong.liu@stevens.edu

Handling Editor: Vijay Mookerjee

Abstract

The global economic disruption brought by COVID-19 crisis can set a stage for the prevalence of financial statement frauds, which jeopardize the efficient functioning of capital markets. In this paper, we propose a nuanced method to detect frauds by tracking granular changes in disclosures over time. Specifically, we first align paragraphs between consecutive disclosures by their similarities. This alignment can be solved as an optimization-based matching problem. Then we identify three types of changed contents: recurrent, newly added, and deleted contents. For each type, we measure the changes in terms of fraud-relevant linguistics features, such as sentiment and uncertainties. Further, we formulate a firm's Management Discussion and Analysis change trajectory over years as a multivariate time series composed of these granular metrics. We implement a deep learning model to predict frauds using the change trajectory as an input. Extensive experiments demonstrate that our model significantly outperforms benchmark models, and its performance increases with the length of the change trajectory. Moreover, we found specific types of changes are strongly associated with frauds, including weak modal or reward words in newly added or deleted contents. Our work provides an optimization-based method to define change trajectories and trace information mutation in narratives. Finally, our study contributes to the fraud detection literature with a new predictive signal—disclosure change trajectories with an effective deep learning architecture.

KEYWORDS
change trajectory, deep learning, financial disclosure fraud, fraud detection, natural language processing

1 | INTRODUCTION

The global economic disruption brought by the COVID-19 crisis has posed unprecedented challenges for firms to meet financial targets and manage stakeholder expectations. Past crises have proven that such a large-scale disruption inevitably sets a stage for the prevalence of financial statement frauds (Anti-Fraud Collaboration, 2021). Corporate financial statements, such as annual 10-K filings, have been an important source of information for the public to understand a firm's operation and potential risks. However, fraudulent statements can pose significant threats to the efficient functioning of capital markets (Dechow et al., 2011). A financial statement fraud (hereafter referred to as fraud for simplicity) is a deliberate attempt by a corporation to deceive or mislead the public by preparing and disseminating a financial statement with material misinformation (Rezaee, 2005). This statement often exaggerates the firm's prospects, manipulates shareholder expectations, or covers the impact of adverse events (Agarwal & Medury, 2014; Huang et al., 2014). Studies have shown that most misstating firms experienced substantial financial problems including dramatic stock price drops, bankruptcies or liquidation, and delisting from the stock market (Beasley et al., 2010; Mahajan et al., 2008). Therefore, such frauds have been serious concerns for investors and regulators such as Securities and Exchange Commission (SEC).

However, identifying frauds is a difficult task because the complexity and length of disclosures have increased, whereas the informativeness has reduced. For example, since the 90s, 10-K filings have become more redundant and contained more boilerplate, hindering their readability and

Accepted by Vijay Mookerjee, after two revisions.

interpretability (Brown & Tucker, 2011; Dyer et al., 2017; Li, 2008). Instead of investigating individual disclosures, studies suggest that disclosures can be better interpreted by closely examining their changes over filing periods (Brown & Tucker, 2011; Cohen et al., 2020). SEC (2003) mandates that the Management Discussion and Analysis (MD&A) section of 10-K filings should "change over time to maintain an appropriate focus on material factors." Cohen et al. (2020) showed that changes in the 10-Ks have strong implications for firms' future returns and operations. Similarly, changes in disclosures can be highly relevant to frauds because frauds are primarily driven by deteriorated financial conditions and firm performance (Rezaee, 2005). However, it has been challenging to pinpoint differences between lengthy disclosures and understand the implications of such changes. As a result, investors are often inattentive to these subtle but powerful signals (Cohen et al., 2020).

Researchers have used machine learning methods to identify frauds, but few have looked into the signals from disclosure changes. Extensive studies have examined the information content of disclosures and attempted to extract informative cues through computational linguistics (Brown & Tucker, 2011; Cecchini et al., 2010a; Dyer et al., 2017; Goel & Uzuner, 2016; Hoberg & Lewis, 2017; Huang et al., 2014; Purda & Skillicorn, 2015). For instance, researchers have analyzed length (Brown & Tucker, 2011), readability (Moffitt & Burns, 2009), tones (Goel & Uzuner, 2016; Loughran & Mcdonald, 2011), and topics (Brown et al., 2020) to identify useful predictors for fraud detection. On the other hand, a stream of financial studies has shown that year-over-year text modification of disclosures is significantly associated with a firm's future earnings, profitability, and critical events such as bankruptcies (Brown & Tucker, 2011; Cohen et al., 2020).

These financial studies offer a new research direction in fraud detection, but many issues remain unexplored. First, it is unclear whether disclosure changes can also be a powerful indicator for frauds. Second, the metrics measuring disclosure changes in these studies offer limited informativeness. These studies calculate the overall modification between two disclosures, for instance, by text similarity or by differences in the total counts of some types of words (Brown & Tucker, 2011; Cohen et al., 2020). However, the increased use of one type of words (e.g., positive words) in one theme can be offset by its decreased use in other themes. These coarse-grained measures may have led to the conclusion that disclosures have incurred fewer changes despite their growing length in the past decades (Dyer et al., 2017; Purda & Skillicorn, 2015). Moreover, the nondirectional text similarity can be ambiguous for fraud detection as firms can make plausible changes (Cohen et al., 2020). In addition, the modification may be caused by new SEC filing requirements or personnel changes, making this signal prone to prediction errors. Finally, extant studies only track how a disclosure differs from the last filing (Brown & Tucker, 2011; Cohen et al., 2020). Notably, severe frauds often span years and can be reflected in multiple disclosures (Beasley et al., 2010; Dechow et al., 2011). This suggests that disclosure changes should be examined within a longer span of filling periods in order to fully capture large-scale frauds.

TABLE 1 Example of change trajectory

Year	10-K excerpts
2004	…our estimates of losses on purchase commitments are based on the assumption that we **will not** receive these conditional price reductions in 2006…
2005	…accordingly, our estimates of our liability for these purchase commitments as of December 31, 2005 are based on the assumption that we **will receive** these conditional price reductions in 2006…
2006	…accordingly, the Company's estimates of its liability for these purchase commitments were adjusted to reflect the fact that the Company **would receive** these conditional price reductions for the remainder of the contract…

To tackle these issues, we propose a new methodology for measuring granular changes in MD&A sections and test whether these measures can effectively predict frauds. We adopt an optimization-based method to overcome the shortcomings of existing change measures. Rather than only considering the overall modification between disclosures, our first task is to align their paragraphs and then examine the detailed changes to the aligned contents. This alignment can be formulated as a maximum weight matching problem (Gerards, 1995), with paragraph similarities as weights. With matched paragraphs, we identify three types of contents: recurrent (maybe modified), new, and deleted contents. Then we measure granular changes in each content type in terms of sentiment, uncertainty, award focus, litigation, and other linguistics features, as suggested by the literature (Cecchini et al., 2010a; Goel & Uzuner, 2016; Hajek & Henriques, 2017; Larcker & Zakolyukina, 2012; Loughran & Mcdonald, 2011). For example, Table 1 shows how a firm discusses a recurrent topic about purchase commitments in its 10-Ks of 2004–2006. Clearly, the highlighted phrases indicate changes in the firm's confidence and uncertainty.

Next, we formulate a change trajectory as a multivariate time series composed of the granular metrics extracted from consecutive disclosures over years. This trajectory can be analyzed to reveal patterns regarding how a firm modified disclosures on a year-over-year basis and may shed light on motivations behind frauds. We then develop a deep learning model to predict frauds using the change trajectory as an input. This model includes a temporal convolution network that selects and configures these granular metrics, and a recurrent neural network that captures their temporal patterns. We tested this model with a dataset of 87,765 disclosures from year 1994 to 2016. We found that this model can achieve an AUC score of 80% and a PRC score of 78%, exceeding benchmark models by a margin of approximately 10% and 14%, respectively.

Our work contributes to the literature in several ways. First, it provides an optimization-based method to define change trajectories in narratives, and this method can be potentially used to track information mutation in other contexts.

Second, we add to the fraud detection literature a new predictive signal: disclosure change trajectories. This signal is not only comprehensive as it captures rich linguistic features, but also interpretable as it allows stakeholders to trace changes to specific business activities over time. Third, our work provides a deep learning model to predict frauds using this signal. We empirically demonstrated that this model can effectively identify the risk of frauds, significantly outperforming benchmark models. As most frauds are related to the results of operations, for example, revenue, goods sold, or inventory (Dechow et al., 2011; Huang et al., 2017), accurately identifying frauds can help pinpoint issues in operations management (OM). Finally, our study generates several new observations regarding financial disclosure frauds. The more MD&A changes, the higher the fraud risk. Frauds are significantly associated with the weak modal words in newly added or removed contents, the negative sentiment of recurrent contents, and the reward focus in newly added or removed contents. These findings contextualize general theories on misinformation and further enhance our understanding of frauds.

2 | THEORETICAL BACKGROUND AND RELATED WORK

A growing body of research uses machine learning and natural language processing (NLP) techniques to detect fraudulent disclosures. As summarized in Appendix 1 of the Supporting Information, researchers have identified two types of useful indicators: (1) quantitative financial ratios (Abbasi et al., 2012; Beneish, 1999; Cecchini et al., 2010b; Craja et al., 2020; Dechow et al., 2011; Kirkos et al. 2007; Lin et al., 2003; Persons, 1995; Zhang et al., 2022), and (2) linguistic and thematic features, including tone, attention focus, emotion, and other psychological behaviors identified from text (Cecchini et al., 2010b; Craja et al., 2020; Goel & Uzuner, 2016; Hajek & Henriques, 2017; Hoberg & Lewis, 2017; Humpherys et al., 2011; Larcker & Zakolyukina, 2012; Loughran & Mcdonald, 2011; Zhang et al., 2022). Researchers have used these indicators to create predictive models for frauds, including Logit Regression, Naïve Bayes, SVM, and deep learning (Brown et al., 2020; Cecchini et al., 2010a; Craja et al., 2020; Dong et al., 2018; Loughran & Mcdonald, 2011; Purda & Skillicorn, 2015).

Previous literature has identified a number of financial ratios that are relevant to frauds, as summarized in Appendix 2 of the Supporting Information. Dechow et al. (2011) created an *F*-score model to flag "wrong-doing" filings with financial indicators such as "Change in Receivables," "Change in Inventory," and "Actual Issuance." *F*-score serves as a benchmark model for fraud detection research. Cecchini et al. (2010b) applied a customized financial kernel with SVM model to detect frauds. Abbasi et al. (2012) proposed a meta-learning model using 12 financial ratios as features and achieved good performance for fraud detection.

Currently, most textual analysis for fraud detection focuses on MD&A sections. SEC requires all firms to discuss the same set of topics within MD&As.[1] In particular, revenue and expenses should be discussed in detail to allow shareholders to observe a firm's performance and operations from the managers' perspectives (Brown & Tucker, 2011; Purda & Skillicorn, 2015). It is also noted that manipulations of revenues and expenses are the basis for most fraud allegations (Hoberg & Lewis, 2017). Hence, studies have found that MD&As are the most relevant in addressing the mistakes in financial statements (Hoberg & Lewis, 2017, Brown & Tucker, 2011; Purda & Skillicorn, 2015).

Extensive studies have been conducted to identify linguistic features from narrative disclosures using well-established dictionaries (Kearney & Liu, 2014). Loughran and Mcdonald (2011) compiled a set of dictionaries (referred to as LM dictionaries) to identify tone, uncertainty, confidence, and other word categories in the financial context. They found that words in "uncertainty," "negative," and "litigious" categories are relevant to frauds (Loughran & Mcdonald, 2011). Another dictionary set, Linguistic Inquiry and Word Count (LIWC; Tausczik & Pennebaker, 2010), has been widely adopted to analyze attention focus, emotionality, thinking styles, and other psychological behaviors from text. Goel and Uzuner (2016) found that sentiment words from both LM and LIWC dictionary sets are highly relevant to frauds. Table 2 summarizes fraud-relevant word categories identified by previous research from these dictionary sets. More details can be found in Appendix 1 of the Supporting Information.

Previous studies (Larcker & Zakolyukina, 2012; Throckmorton et al., 2015) also indicate that most of the word categories in Table 2 are well-aligned with deception theories in four perspectives: emotion, lack of embracement, cognitive effort, and attempted control (Vrij, 2008). The emotion perspective suggests that because deceivers are afraid to be caught, they might experience negative emotions as manifested in their negative statements. Thus, sentiment words (positive or negative) in a statement are generally recognized as good predictors for deceptions (Goel & Uzuner, 2016; Loughran & Mcdonald, 2011). The lack of embracement perspective argues that deceivers lack conviction and differ from truth-tellers on the degree of certainty in their statements. Words in the uncertainty, weak modal, and strong modal categories have been widely used to measure narrative uncertainties (e.g., Loughran & Mcdonald, 2011; Larcker & Zakolyukina, 2012). The cognitive effort perspective implies that deceptive statements are likely to lack concrete details because it is difficult to fabricate such details coherently. Similarly, the control perspective suggests that deceivers prefer to use general nonspecific languages, short statements without details, and few self-references so that they can control deceptive contents to avoid self-incriminating. These theoretical perspectives can explain why words in the reward and achievement categories can help detect frauds. As these words describe motives (i.e., opportunity, gain, win, etc.) that drive or guide a person to behave (Pennebaker et al., 2015), they often imply positive emotions, self-references, and concrete concepts. Previous studies have found that these words are positively associated with language persuasiveness (Xiao,

TABLE 2 Word categories relevant to frauds

Word categories	Description
Negative (Neg)	Negative words in the financial context, such as "abnormal," "abuse," "broken."
Positive (Pos)	Positive words in the financial context, such as "effective," "good," "improve."
Uncertainty	Words denoting uncertainty and imprecision in the financial context, such as "approximate," "believe," "confuse."
Litigious	Words that are relevant to the legal contest, litigiousness, or litigious environment in the financial context, such as "allege," "amend," "restate."
Strong modal (SM)	Words reflecting strong or high levels of confidence in the financial context, such as "must," "never," "always."
Weak modal (WM)	Words reflecting weak or low levels of confidence, or high levels of uncertainties in the financial context, such as "possible," "may," "depend." The Weak Modal category is a subset of the Uncertainty category.
Comparatives (Compare)	Words involving a comparison expression from a psychological perspective, such as "greater," "best," and "worse."
Reward	Words referencing to rewards, incentives, and positive goals, such as "achieve," "promote," "success."
Achievement (Achieve)	Words indicating a result gained by effort, such as "ambition," "beat," "honor."
Discrepancy (Discrep)	Words such as "expect," "hope," "need," "should."

2018), but negatively correlated with troll tweets (Addawood et al., 2019) and misinformation (Clarke et al., 2021; Jiang & Wilson, 2018).

In addition, there are a number of studies using topic modeling to identify topics that can differentiate fraudulent disclosures from truthful ones (Brown et al., 2020; Dong et al., 2018; Hoberg & Lewis, 2017). Hoberg and Lewis (2017) compared fraudulent with nonfraudulent MD&As, and found fraudulent firms likely grandstand good performance with few details disclosed. Interestingly, Dong et al. (2018) extracted signals such as sentiment, emotion, topics, and social network features from social media, and then combined these signals with MD&As and financial indicators to predict frauds.

Despite a rich set of proposed textual features, researchers also found that these features have become less effective in analyzing individual disclosures because firms learned to deliberately avoid words (e.g., negative words) defined in the well-accepted dictionaries (Cao et al., 2020). Moreover, MD&As tend to become more similar with fewer modifications over time (Brown & Tucker, 2011). Nevertheless, studies found that the changes between a firm's consecutive filings can provide strong predictive signals (Brown & Tucker, 2011; Cohen et al., 2020). Cohen et al. (2020) assembled a dataset of filings to measure the quarter-over-quarter text similarity as well as sentiment changes. Their study shows that the modifications in the 10-Ks can predict future earnings, stock price, profitability, or bankruptcies. Taking a different approach, Purda and Skillicorn (2015) first computed a metric called probability-of-truth for individual disclosures and then showed that the change of this metric from one quarter to the next had an incremental predictive power in identifying frauds. In a similar vein, a recent study finds that executives' tone changes in consecutive earnings calls are strongly associated with stock returns (Druz et al., 2020).

In domains other than finance, text analytics has been widely used to analyze business documents or user generated contents accumulated over time. In OM, customer reviews have been used to assist product defect discovery (Abrahams et al., 2015), forecast sales (Chong et al., 2016; Lau et al., 2018), and infer operational efficiency (Ko et al., 2019). As customer opinions change over time, tracing the changes on various sentiment aspects (e.g., quality, price) may shed more insights on the dynamics of operations. For product family design (Jiao et al., 2007), text analytics is applied to discover similarities among product variants based on product specifications (Jiao et al., 2007). As customization leads to high product variety, framing these variants into change trajectories can pinpoint their commonalities and differences to support product family design. Finally, information mutation during the diffusion process can also be modeled as change trajectories. Moussaïd et al. (2015) studied how messages regarding public hazard events undergo changes when passed through subjects. They found the messages become shorter, gradually inaccurate, and increasingly dissimilar, while the perception of risk is amplified through the diffusion. Shin et al. (2018) studied the mutation of rumors and hypothesized that the need for changes is saliently present for resurging rumors.

Motivated by these studies, we would like to explore the effectiveness of MD&A changes as a signal toward fraud detection. However, several challenges remain. First, these studies only measure the overall similarity without locating where specific changes have happened from one disclosure to another. We attempt to pinpoint specific changes to picture how a fraud unfolds and to provide better interpretation of predictions. Moreover, the overall similarity cannot tell whether an MD&A underwent plausible or abnormal changes. Second, these studies usually measure the changes in the total count of the words in a category (e.g., sentiment words). However, if the words are increased in one paragraph but reduced in another, their total count almost remains the same. In addition, existing studies only consider changes within two consecutive filing periods. These short-term changes may be caused by new filing requirements mandated by SEC, changes in executive teams, or other reasons irrelevant to frauds (Brown & Tucker, 2011; Cohen et al., 2020). Finally, about one third of frauds last for at least three filling periods,[2] and the average duration of a fraud is

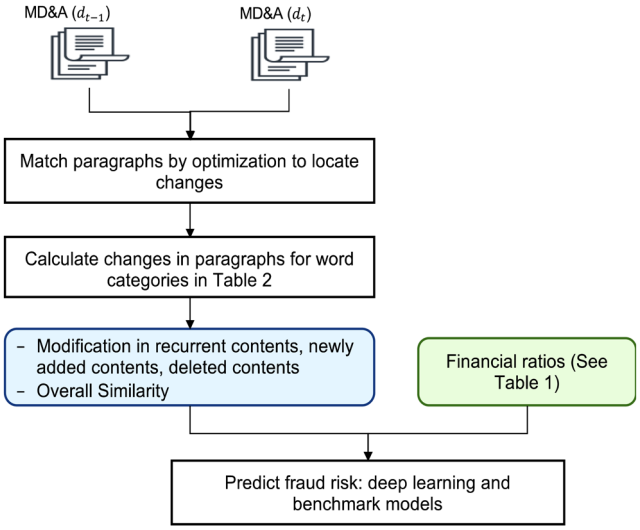

FIGURE 1 Overall architecture [Color figure can be viewed at wileyonlinelibrary.com]

around 31.4 months (Beasley et al., 2010). Thus, it would be desirable to consider a firm's MD&A change trajectory over a longer period to effectively and timely detect continuous frauds.

3 | METHODOLOGY

To overcome these challenges, we proposed a new method to measure granular MD&A changes and change trajectories. The overall architecture is shown in Figure 1. We first align paragraphs in two consecutive MD&As using a weighted matching technique. After matching, paragraphs can be categorized into three types: (1) *recurrent paragraphs*, that is, a pair of matched paragraphs discuss the same topic, but differ in wording, (2) *newly added paragraphs*, and (3) *paragraphs deleted* from the previous MD&A. For each type, we measure the content in terms of sentiment, uncertainty, and other aspects as listed in Table 2.

Then, we use these granular change measures along with financial ratios to build models to predict frauds. We create a deep learning model and benchmark its performance with a number of machine learning models that have been well-adopted by previous fraud detecting research (Abbasi et al., 2012; Goel & Uzuner, 2016; Humpherys et al., 2011; Larcker & Zakolyukina, 2012; Purda & Skillicorn, 2015), as well as models in multivariant time-series regression (Ruiz et al., 2021). Next, we describe each component, starting with how to determine MD&A changes.

3.1 | Matching paragraphs to locate changes

Different from previous work that just measures the overall changes between two consecutive MD&As (Brown and Tucker, 2011; Cohen et al., 2020), our objective is to locate which specific contents have been modified since the last MD&A. The most granular changes can happen at the paragraph level because a paragraph in MD&A usually deals with a single theme (Dyer et al., 2017). Thus, we segment each MD&A into paragraphs and preprocess them by removing tables, punctuations, and numbers. We also remove company names from MD&As because they are frequent and can inflate paragraph similarities. In addition, we discard paragraphs with less than 20 words because these paragraphs are usually headings.

TABLE 3 Paragraph matching example

	$p_{1(t)}$	$p_{2(t)}$	$p_{3(t)}$	$p_{4(t)}$	$p_{5(t)}$
$p_{1(t-1)}$	**0.8**	0.5	0.6	0.4	0.2
$p_{2(t-1)}$	0.6	0.3	**0.9**	0.4	0.1
$p_{3(t-1)}$	0.1	**0.7**	0.3	0.2	0.1
$p_{4(t-1)}$	0.3	0.2	0.2	0.3	0.1
$p_{5(t-1)}$	0.1	0.1	0.3	**0.8**	0.2
$p_{6(t-1)}$	0.4	0.3	0.3	0.2	**0.7**

The optimal match is shown as bold values.

We measure the granular changes between two consecutive MD&As at years t and $t-1$ (denoted as d_t and d_{t-1}, respectively) as follows. With all the paragraphs forming a corpus, we represent each paragraph by its TF-IDF (term frequency/inverse document frequency) vector. It has been well-accepted that a document in a corpus can be represented as a vector of TF-IDF weights (Schütze et al., 2008). Alternatively, each paragraph can also be embedded into a vector called BERT embedding using the latest sentence transformer model called SBERT (Reimers & Gurevych, 2019). We will first use TF-IDF embeddings to illustrate our methodology from end-to-end, and then switch to BERT embeddings. The impact of embedding techniques on the model performance will be discussed later.

Next, we calculate the pairwise cosine similarities between paragraphs in d_t and d_{t-1}. Then, with paragraph similarities, we match paragraphs in d_t with those in d_{t-1} such that their overall similarity is maximized. Formally, let $s_{i,j}$ be the cosine similarity between paragraphs i and j, and $x_{i,j}$ be a binary variable indicating whether paragraph i is matched with j, where $i \in d_t$ and $j \in d_{t-1}$, the matching can be formulated as an optimization problem:

$$\begin{aligned} &\text{Objective}: \quad \max \textstyle\sum_{i\in d_t} \sum_{j\in d_{t-1}} x_{i,j} s_{i,j}, \\ &\text{subject to} \quad x_{i,j} = 0 \text{ or } 1, \\ &\qquad\qquad 0 \le \textstyle\sum_{i\in d_t} x_{i,j} \le 1, 0 \le \sum_{j\in d_{t-1}} x_{i,j} \le 1. \end{aligned} \tag{1}$$

To illustrate this problem, let us consider similarities (s_{ij}) between paragraphs in d_t and d_{t-1} shown in Table 3. We wish to assign each paragraph in d_t to at most one paragraph in d_{t-1} such that the sum of the similarities of the matched pairs is maximized. This optimization problem can be solved using the Kuhn–Munkres algorithm (Munkres, 1957). An outline

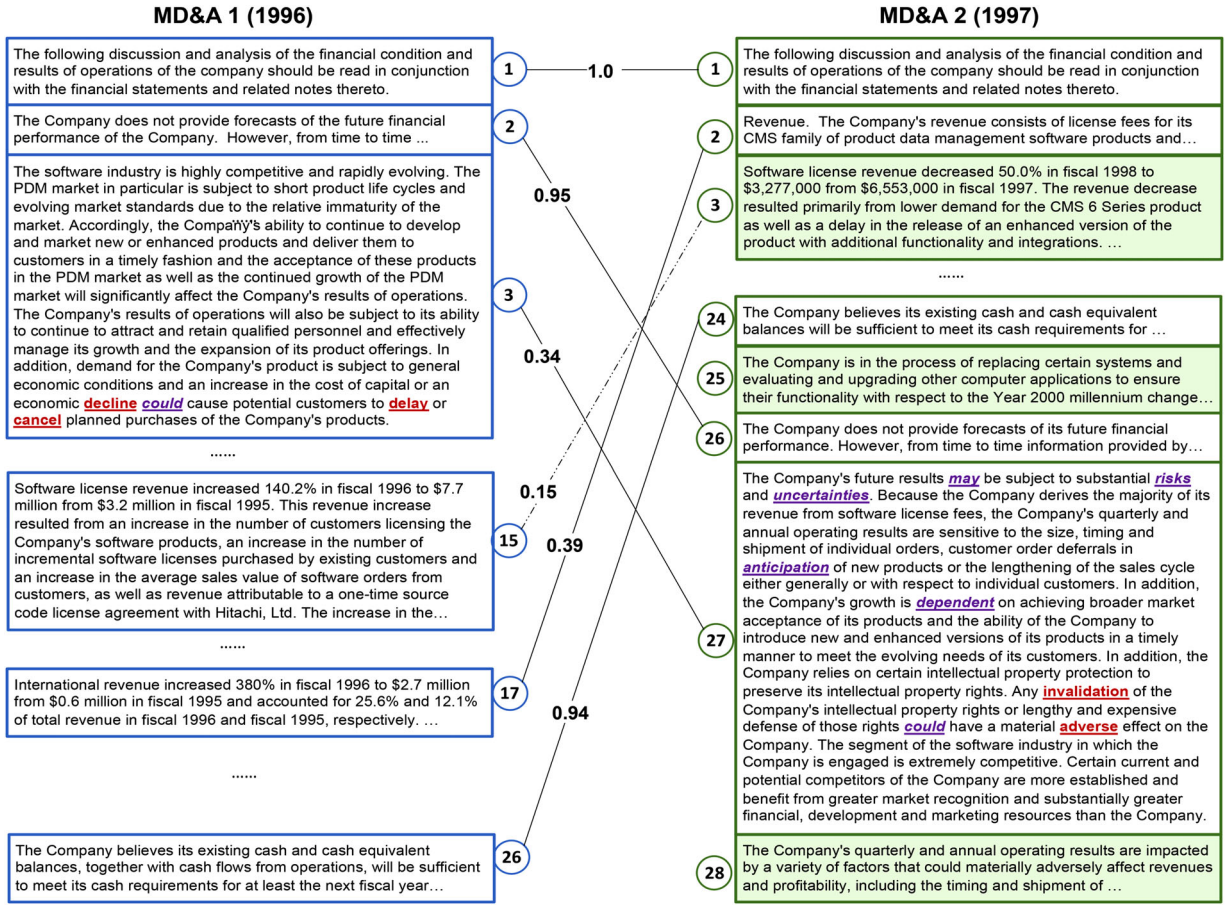

FIGURE 2 Matching paragraphs in two consecutive MD&As (best viewed in color) [Color figure can be viewed at wileyonlinelibrary.com]

of the algorithm can be found in Appendix 3 of the Supporting Information. The optimal assignment for this example is highlighted in bold in Table 3. For example, paragraph $p_{1(t)}$ in d_t is matched with $p_{1(t-1)}$ in d_{t-1}. However, $p_{1(t)}$ has been modified from $p_{1(t-1)}$,as their similarity is 0.8. Also, note that d_{t-1} has one more paragraph than d_t. As a result, $p_{4(t-1)}$ in d_{t-1} cannot be matched to any paragraph in d_t. Thus, $p_{4(t-1)}$ represents some content deleted from d_{t-1}. On the contrary, if a paragraph, say $p_{i(t)}$ in d_t, cannot be matched with any paragraph in d_{t-1}, then $p_{i(t)}$ is the new content added in d_t.[3]

Sometimes, although two paragraphs are matched, their similarity score is fairly low. Very likely, these paragraphs discuss different themes. We set a threshold *th*, and modify Equation (1) to require $s_{i,j} \geq th$ when $x_{i,j} = 1$. Only if the similarity of two paragraphs is greater than *th*, they can be matched. To determine *th*, we randomly selected 100 pairs of paragraphs with evenly distributed similarity scores, and asked two graduate students with finance backgrounds to manually annotate whether each pair concerns the same theme. Among the pairs with similarities higher than 0.30, 97% of them were labeled to have the same themes by both annotators.[4] Thus, we set $th = 0.3$.

To illustrate the necessity of matching, in Figure 2, we show the alignment results of two successive MD&As[5] (denoted as MD&A_1996 and MD&A_1997) through TF-IDF embeddings. After calculating pairwise paragraph similarities, we found that most paragraphs in MD&A_1997 are recurrent contents but have been reordered from MD&A_1996. For instance, the second paragraph in MD&A_1996 is mapped to Paragraph 26 at the end of MD&A_1997, with a high similarity score of 0.95. Another pair of matched paragraphs, Paragraph 15 in MD&A_1996 and Paragraph 3 in MD&A_1997, discuss software license revenue, but one focuses on the revenue increase while the other describes product license fees. As the similarity score of this pair is as low as 0.15, Paragraph 15 is considered as a deleted content from MD&A_1996 (highlighted in light gray) and Paragraph 3 is treated as a newly added content (highlighted in light green). Paragraphs 25 and 28 in MD&A_1997, covering an emerging topic Y2K and some risk discussion, cannot be mapped to any paragraphs in MD&A_1996. Hence, they are also classified as newly added paragraphs.

3.2 | Measuring granular changes in paragraphs

With matched paragraphs, now we can calculate overall similarities using Equation (1). We use the mean similarity over

TABLE 4 Metrics measuring granular MD&A changes

Metric	Description
Overall similarity between d_t and $d_{t-1}(s_t)$	The sum of the similarities of matched paragraph pairs divided by the maximum paragraph numbers of d_t and d_{t-1}, that is, $s_t = \frac{\sum_{i\in d_t}\sum_{j\in d_{t-1}} x_{i,j}\ s_{i,j}}{\max\{\lvert d_{t-1}\rvert,\lvert d_t\rvert\}}$, for example, for Table 3, $s_t = \frac{(0.8+0.7+0.9+0.8+0.7)}{6} = 0.65$
Granular word change metrics for category C, $C \in \mathbb{C}$	
add_t^C	The average percentage of C-words in the paragraphs newly added to d_t
del_t^C	The average percentage of C-words in the paragraphs that exist in d_{t-1} but not in d_t
$up_t^C(down_t^C)$	The percentage of recurrent paragraphs in d_t with upticks (downticks) in C-words.
MD&A change trajectory in the past T years	
$Traj_t(T) = \{x_{t-T+1}, \dots, x_{t-1}, x_t\}$, $x_i = \{s_i, [add_i^C, del_i^C, up_i^C, down_i^C]^{C\in\mathbb{C}}\}$, for $i \in [t-T+1, t]$	

the maximum number of paragraphs to measure the overall similarity between d_t and d_{t-1} (denoted as s_t). Note that s_t allows us to determine how much an MD&A deviates from the firm's reporting routine. For example, for the MD&As in Table 3, $s_t = 0.65$, because the total similarity of the matched pairs is 3.9, and the maximum number of paragraphs is 6.

Moreover, previous studies have identified a number of word categories (denoted as $\mathbb{C}$) that are relevant to frauds, as summarized in Table 2. With aligned paragraphs, we calculate four types of *granular changes* for words of each category C (denoted as C-words): *uptick or downtick, newly added*, or *deleted words*, as shown in Table 4. Take the category "Uncertainty" (i.e., $C = Uncert$) as an example, assuming that five paragraphs in d_t have no counterparts in d_{t-1}, we first calculate the percentage of uncertainty words in each of the paragraphs and then take the average of these five percentages as add_t^{Uncert}. Intuitively, if newly added paragraphs frequently use uncertainty words, it may indicate that the firm encountered some worrisome new conditions. For paragraphs in d_{t-1} but without counterparts in d_t, del_t^{Uncert} can be calculated similarly.

For recurrent paragraphs, we measure up_t^C ($down_t^C$) as the percentage of paragraphs in d_t that contain 10% more (or less) C-words than the counterparts in d_{t-1}. We consider changes below 10% are insignificant. For instance, for the aligned paragraph pair 3–27 shown in Figure 2, there is only one uncertainty word (highlighted in underlined italics) in Paragraph 3, while Paragraph 27 has six uncertainty words, increased by five times. Therefore, an uptick in uncertainty words can be found in Paragraph 27. Similarly, compared to Paragraph 3, Paragraph 27 has one third fewer negative words (highlighted in underlined bold). In total, regarding negative words, out of the 24 recurrent paragraphs in the two MD&As, 2 paragraphs have downticks, 2 paragraphs see upticks, and 20 remain unchanged. Therefore, up_t^{Neg} and $down_t^{Neg}$ are both 0.083 (i.e., 2/24).

Next, we define change trajectories. For each MD&A (d_t) and each word category C, we can calculate four granular word change metrics for d_t, as shown in Table 4. With 10 categories of words, in total, we use 40 metrics along with the overall similarity (s_t) to measure the changes in d_t from d_{t-1}. Therefore, a firm's MD&A *change trajectory in the past T years* can be represented as a time series:

$Traject(T) = \{x_{t-T+1}, \dots, x_{t-1}, x_t\}$, where

$$x_i = \Big\{ s_i, \Big[add_i^{C_1}, del_i^{C_1}, up_i^{C_1}, down_i^{C_1} \Big], \times \Big[add_i^{C_2}, del_i^{C_2}, up_i^{C_2}, down_i^{C_2} \Big], \dots, \times \Big[add_i^{C_{10}}, del_i^{C_{10}}, up_i^{C_{10}}, down_i^{C_{10}} \Big] \Big\}, \text{ for } i \in [t-T+1, t]. \quad (2)$$

3.3 | Comparing our change measurement with existing methods

In general, previous studies measure changes in two consecutive MD&As using the following metrics:

- *Modification score* that tracks the number of edits (inserted or deleted words) required in order to change one document into the other (Cohen et al., 2020). Then, changes for each word category, say C, can be measured as the percentages of C-words inserted or deleted out of the total edits, denoted as $C+$ or $C-$, respectively.
- *Cosine similarity* of the TF-IDF vectors representing two documents (Brown and Tucker, 2011; Cohen et al., 2020).
- *Jaccard score* calculated as the ratio of shared words to the union of all the words in these two documents (Cohen et al., 2020).
- Topic mixture changes that measure the differences in thematic topics discovered from MD&As by topic modeling (Brown et al., 2020; Dyer et al., 2017).

Our granular metrics differ from these metrics in several aspects. First, as we align paragraphs before measuring, our change metrics are more accurate than the modification score. For instance, the two MD&As in Figure 2 have mostly similar but reordered paragraphs. Without alignment, in order to change MD&A_1996 into MD&A_1997, massive edits must

be made starting from Paragraph 2. As a result, the modification score is 0.92, that is, 92% of words inserted or deleted out of the total number of words in these two documents. In contrast, with alignment, the overall similarity (i.e., s_t as defined in Table 4) between them is 0.76. In other words, their difference is only 0.24, dramatically lower than 0.92.

Second, compared with Cosine similarity or Jaccard score, our method can locate specific changes and capture the direction of changes in words. Cosine similarity or Jaccard score only measures the overall similarity without showing upticks or downticks in specific words. Moreover, such measures may not capture meaningful changes because the addition of words in one paragraph can be offset by the removal of such words in other paragraphs.

Finally, studies have proposed to measure the differences in topic mixtures of MD&As. This method requires fitting topic models, as well as manual interpretation of discovered topics. Usually, only mainstream issues prevailing in most MD&As can be discovered as topics, while issues very specific to a firm rarely emerge as topics. Hence, changes identified by this method may lack fine granularity. To summarize, our method can effectively address the drawbacks in the extant work. By paragraph matching, our method can precisely locate changes; and through a set of well-calibrated metrics, we can accurately capture both overall similarity and granular changes between consecutive MD&As.

3.4 | Prediction models

Our prediction task can be formulated as follows. Let $fraud_t^k$ indicate whether the filing of firm k in year t is fraudulent. $fraud_t^k = 1$, if it is a fraud, and 0 otherwise. Our independent variables contain financial ratios (see Appendix 2 of the Supporting Information) and change trajectories in T years (see Equation 2). Our target is to predict $p\ (fraud_t^k = 1 \mid Fiancial\ Ratios_t^k(T), Trajectory_t^k(T))$.

Deep learning models have been widely used in time-series analysis due to their superior capabilities in capturing complex relationships in variables. We carefully crafted a deep learning model to predict frauds as shown in Figure 3. The model has three inputs. The first input contains 40 features of word changes in total. We include a special feature component based on temporal convolutional network (TCN; Lea et al., 2017) to select and configure word change features. More details of TCN will be given shortly. Then, to capture the patterns in MD&As change trajectories, the configured word change features, concatenated with the overall similarity and financial indicators, enter a long short-term memory (LSTM) unit (Hochreiter & Schmidhuber, 1995), a special type of recurrent neural networks. LSTM is a well-accepted structure for extracting features from sequence inputs, such as time-series data. We adopt a bidirectional LSTM (BiLSTM) layer so that the features for any year i are encoded with references to features before and after year i. This encoding is sent to a fully connected layer to produce the final prediction of fraud risk.

Now we describe TCN, the important feature selection component of our model. Our word change metrics have complex relationships. For instance, when discussing a new concerning condition, a firm may use more negative words (add_t^{Neg}) and less positive words (add_t^{Pos}) in newly added paragraphs. If this condition adversely affected existing operations, an increasingly negative tone can also be seen in the recurrent paragraphs ($up_t^{Neg}, down_t^{Pos}$). Therefore, it would be desirable to automatically select and transform the most responsive word change metrics to capture effective signals of frauds. We choose TCN (Lea et al., 2017), a special type of convolution neural network, to achieve this task. A TCN has several interesting characteristics. First, a TCN layer produces an output sequence in the same length as the input sequence. Second, by using zero padding, an output element at timestep t, say o_t, only depends on input elements $\{x_1, x_2, \ldots, x_t\}$, but not on those after t. In other words, an output element is selectively composed of a subset of the input elements.

As shown in Figure 3, we first prepend the input of raw word change metrics (in the shape of 10×4) with zero vectors to enlarge its shape to 19×4. Let the enlarged metrics be $\boldsymbol{x'} = \{0, 0, \ldots, 0, x_1, x_2, \ldots, x_{10}\}$, where each x contains four metrics for a word category, $\boldsymbol{x'} \in R^{19\times4}$. Then we apply a filter, that is, a tunable parameter matrix $\boldsymbol{w} = \{w_1, w_2, \ldots, w_{10}\}$, where $\boldsymbol{w} \in R^{10\times4}$, to $\boldsymbol{x'}$ to produce a vector of 10 convoluted features $\boldsymbol{z} = \{z_1, z_2, \ldots, z_{10}\}$, and each feature z_j is calculated as:

$$z_j = ReLU\left(\sum_{k=1}^{10} w_k^T x'_{k+j-1}\right), \text{for} 1 \le j \le 10,$$
$$\text{where activation function } ReLU(u) = \max(0, u). \tag{3}$$

In other words, each convoluted feature z is the sum of the elementwise product of $\boldsymbol{w}$ and a region of $\boldsymbol{x'}$ in the same size as $\boldsymbol{w}$. To illustrate,

z_1: the sum of the elementwise product of $\boldsymbol{w}$ and $\{0, 0, 0, 0, 0, 0, 0, 0, 0, x_1\}$,
z_2: the sum of the elementwise product of $\boldsymbol{w}$ and $\{0, 0, 0, 0, 0, 0, 0, 0, x_1, x_2\}, \ldots$, and
z_{10} : the sum of the elementwise product of $\boldsymbol{w}$ and $\{x_1, x_2, \ldots, x_{10}\}$.

Because of padding, the first convoluted feature only depends on the first word category (i.e., weak modal words as shown in Figure 3), the second on the first two categories (i.e., both weak modal and uncertainty words), and so on, until the last one is composed of all word change metrics. We organize the list of word categories by their correlations with the target variable $fraud_t^k$, such that significant word categories participate more in the output elements.[6] To ensure features can effectively respond to different fraud patterns, we use F filters, each producing one convoluted feature vector. F is

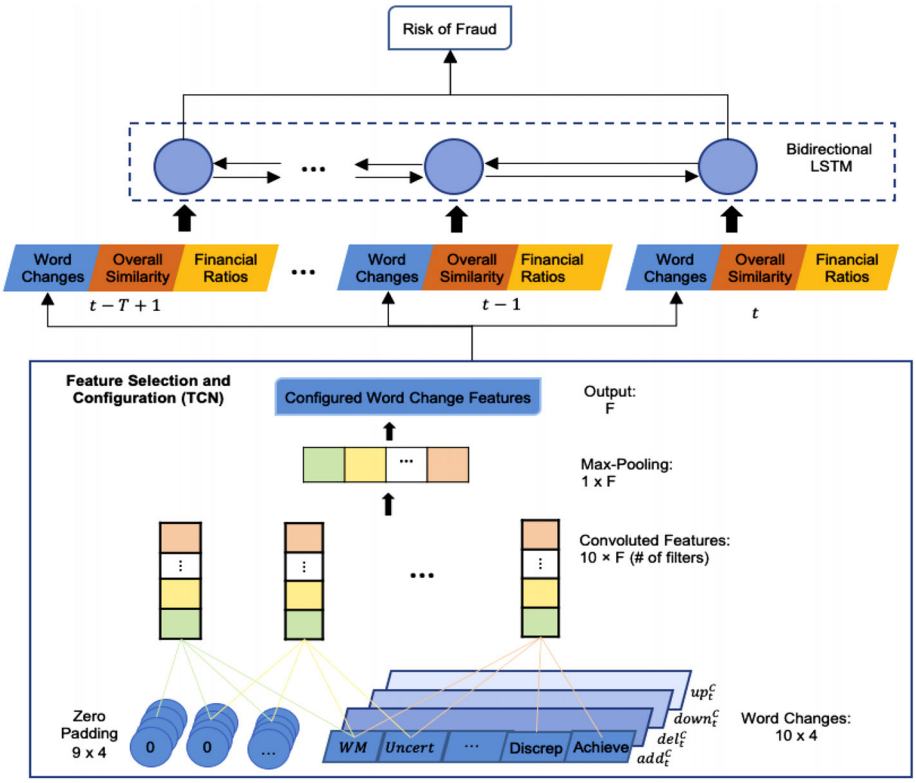

FIGURE 3 Architecture of deep learning model [Color figure can be viewed at wileyonlinelibrary.com]

a hyperparameter tuned during model training. Then a Max Pooling layer is applied to select the largest element from each convoluted vector. The final output of TCN consists of F configured word features. These features join the overall MD&A similarity and financial ratios to enter the BiLSTM unit as shown in Figure 3.

4 | EXPERIMENTS AND RESULTS

Next, we implement our methodology and test whether change trajectories can provide powerful predictability for frauds. We first describe our dataset.

4.1 | Datasets

We obtained a manually compiled dataset of Accounting and Auditing Enforcement Releases (AAERs),[7] which is widely used in fraud analysis (Abbasi et al., 2012; Brown et al., 2020; Dechow et al., 2011). Out of 4012 SEC AAERs issued between 1994 and 2018, the dataset confirms 886 10-K frauds with the last one occurring in 2016. To match this dataset, we retrieved 251,070 10-K documents filed during the same time period from SEC's EDGAR system. Following previous work on fraud detection (Abbasi et al., 2012; Cole & Jones, 2005), we removed disclosures filed by firms in utility, bank, and insurance industries for the sake of consistency, because SEC has different disclosure guidance for these industries. Then we successfully obtained 151,539 MD&As after excluding those that cannot be parsed due to irregular formats. We also retrieved the financial ratios shown in Appendix 2 of the Supporting Information from COMPUSTAT. After matching MD&A with COMPUSTAT data,[8] we have a comprehensive final dataset consisting of 87,765 firm-year observations for 11,303 firms with 720 confirmed frauds spanning from 1994 to 2016.[9]

Table 4A-1 in Appendix 4 of the Supporting Information describes the preprocessing steps and the sample size after each step.

We track how a firm's MD&A filed in each year was changed from its previous filing. Table 5 shows the distribution of frauds and observations for each year. Starting from 1995, the number of frauds increased gradually, culminated in 2001, and then decreased.[10]

Note that our dataset is extremely imbalanced. There are only 720 (0.82%) frauds out of 87,765 firm-year observations.

In addition, as noted by Habib and Hossain (2013), executive personnel changes may lead to dramatic modification of financial reports. To rule out the confounding effects of such changes, following the work by Cohen et al. (2020), we introduce a dummy variable (*Executive_Change*) to denote whether a disclosure mentions executive changes and scan 10-Ks for such clauses.[11]

We found 13% 10-Ks mention executive changes, and as expected, *Executive_Change* has a positive correlation with MD&A overall change ($1 - s_t$), but the coefficient is only

TABLE 5 Observation distribution by year

Year	Firm-year observations	Frauds	Fraud %	Year	Firm-year observations	Frauds	Fraud %
1995	1277	8	0.63	2007	4030	25	0.62
1996	2759	18	0.65	2008	3839	19	0.49
1997	5093	43	0.84	2009	3698	20	0.54
1998	5182	49	0.95	2010	3546	20	0.56
1999	5067	63	1.24	2011	3456	17	0.49
2000	5079	69	1.36	2012	3426	23	0.67
2001	5085	75	1.47	2013	3426	11	0.32
2002	4885	68	1.39	2014	3477	5	0.14
2003	4645	63	1.36	2015	3448	1	0.03
2004	4507	54	1.20	2016	3301	0	0.00
2005	4337	41	0.95	Total	87,765	720	0.82
2006	4202	28	0.67				

0.032 (p *value* < 0.01). During experiments, we concatenate Executive_*Change* variable with s_t and considered it as a part of the MD&A overall change.

More details of our dataset can be found in Appendix 4 of the Supporting Information. On average, each MD&A has 5348 words and 60 paragraphs. The average dissimilarity (or overall changes) is 0.38. This is consistent with the manual analysis about MD&A changes by Brown and Tucker (2011) which confirms that averagely MD&As have about 70% repeated aspects (or topics) and 30% different aspects. We take Weak Modal as an example. On average, about 3% of recurrent paragraphs have upticks or downticks in weak modal words, and 0.29% (0.27%) of words in newly added (deleted) paragraphs are weak modal words. In contrast, without aligning paragraphs, the modification score calculated by the method proposed by Cohen et al. (2020) claims an average of 46% overall changes in MD&As, which over-calculates the changes by 7%, compared to our method.

4.2 | Model training and results

We benchmark our deep learning model with Logistics Regression, SVM, XGBoost, and Random Forest models, which have been well-adopted by previous fraud detecting research (Abbasi et al., 2012; Goel & Uzuner, 2016; Humpherys et al., 2011; Larcker & Zakolyukina, 2012; Purda & Skillicorn, 2015). In addition, we include two recent models for time-series classification: LSTM and Hierarchical Vote Collective of Transformation-based Ensembles V2.0 (HIVE-COTE; Ruiz et al., 2021). HIVE-COTE is a heterogeneous meta ensemble from a shapelet (i.e., representative subsequences) model, a tree-based classifier, a CNN-based neural network, and others. It is considered as the state-of-the-art model in multivariate time-series regression (Ruiz et al., 2021). We train all models using two strategies:

- *Traditional fourfold cross-validation*: To compare model performance, we conduct traditional cross-validation and report the average performance of the test subsets. This strategy allows us to make full use of all positive observations to create a relatively rich dataset for model training and testing. With 720 frauds in total, we employ fourfold cross-validation to ensure sufficient positive observations in each test set.
- *Walk forward validation*: We adopt walk forward validation to test how our model performs in a more realistic setup. We use firm-year observations before a specific year, say t, to train a model and then report the out-of-sample performance of the observations for year t.

As our fraud class is severely underrepresented (<1%), we take two widely used strategies targeting at imbalanced datasets: sampling and cost-sensitive learning (He & Garcia, 2009). In the first strategy, we randomly sample the same number of negative observations as the fraud cases to form a training dataset. To overcome the deficiency of information loss introduced by undersampling, we random sample 100 subsets of negative observations and train 100 models to form an easy ensemble (He & Garcia, 2009). Then we report the average test performance out of the 100 models. In line with existing work (Abbasi et al., 2012; Craja et al., 2020), we use AUC to measure the overall performance of the models. We report precision, recall, and $F-1$ scores at the default threshold 0.5. Meanwhile, we calculate precision and recall under different thresholds to obtain the precision-recall curve (PRC) and report the area under it as another overall performance metric. For the cost-sensitive learning strategy, we keep all training samples but associate a higher cost for misclassifying positive samples. As both strategies provide comparable performance in terms of the overall AUC performance, we present the results obtained by cost-sensitive learning in Appendix 5 of the Supporting Information. The details of implementation and parameter tuning of our deep

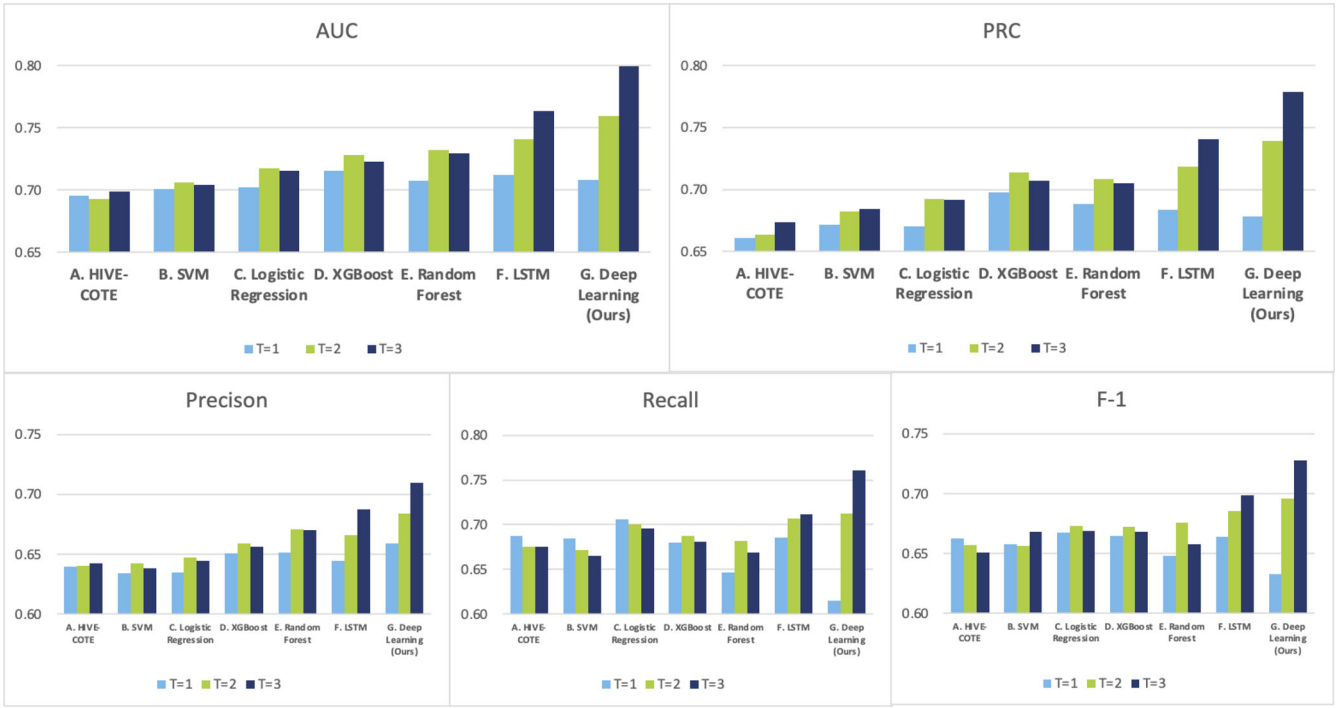

FIGURE 4 Comparing model performance by cross-validation [Color figure can be viewed at wileyonlinelibrary.com]

learning model can be found in Appendix 6 of the Supporting Information.

4.2.1 | Cross-validation results

As shown in Figure 4, we first compare our deep learning model with six baseline models (A–F). All the models are trained with the full inputs, including financial ratios and change trajectories (see Figure 3). As models B–E cannot encode time series, we concatenate these input metrics over T years into a list and feed the list as model inputs.

When $T = 1$, that is, only changes in MD&As of year t from year $t-1$ are considered, all models show similar performance, with AUC around 0.70, and precision, recall, and PRC all around 0.65. However, when T increases, the advantage of the deep learning model (Model G) becomes apparent. When $T = 2$, the AUC is increased by 5%, PRC by 6%, precision by 2%, and recall by 9%. All the metrics culminate at $T = 3$. Compared with the result at $T = 1$, Model G achieves an AUC score of 80%, 9% higher, and the PRC score reaches 78% at $T = 3$, increased by 10%. Similar increase can be observed from the precision and recall scores. In particular, when $T = 3$, our model on average can retrieve 76% fraudulent cases with a precision score of 71%. This experiment demonstrates that with our deep learning model, a longer change trajectory can provide more powerful indication for frauds.

However, with regard to the baseline models A–F, the increase of T has mixed effects on their performance. Surprisingly, HIVE-COTE as a designated multivariate time-series model underperforms in our case, largely because it is designed to detect patterns or shapelets from relatively long sequences (e.g., $T \geq 10$), while our time series is rather short. For non-time-series models B–E, the AUC scores of Models A–D gather around 0.70 and the PRC scores rise slightly above 0.65 when $T = 3$. The Random Forest model (Model E) performs the best in this group, but its performance declines when $T = 3$. This suggests that these non-time-series models are unable to utilize the signals encoded in the change trajectory. On the contrary, the LSTM (Model F) performs the best among all the baseline models, because it can capture the temporal dependencies encoded in the short multivariate time series. In fact, the LSTM model resembles our model G except that it has a LSTM layer instead of a BiLSTM and does not use a TCN module (see Figure 3). Next, we conduct a series of ablation analyses to understand the contribution of each component in the deep learning model G.

Ablation on input components: To understand the importance of each model input (see Figure 3), we create Models H–L from Model G by removing input components, as shown in Figure 5. First, Model H is trained with the change trajectory only (i.e., the financial ratios are removed from its inputs). As T increases, again, we can observe significant upticks in all the four metrics. However, compared with the full Model G, these metrics drop about 5%–11% across all T values. Model I is trained with the financial ratios only (i.e., the change trajectory is removed from Model G). This model receives reduced AUC and PRC scores, about 7% lower than Model G, when $T = 3$. This comparison indicates that the change trajectory plays a critical role in the prediction. Further, to estimate the impact of the granular word

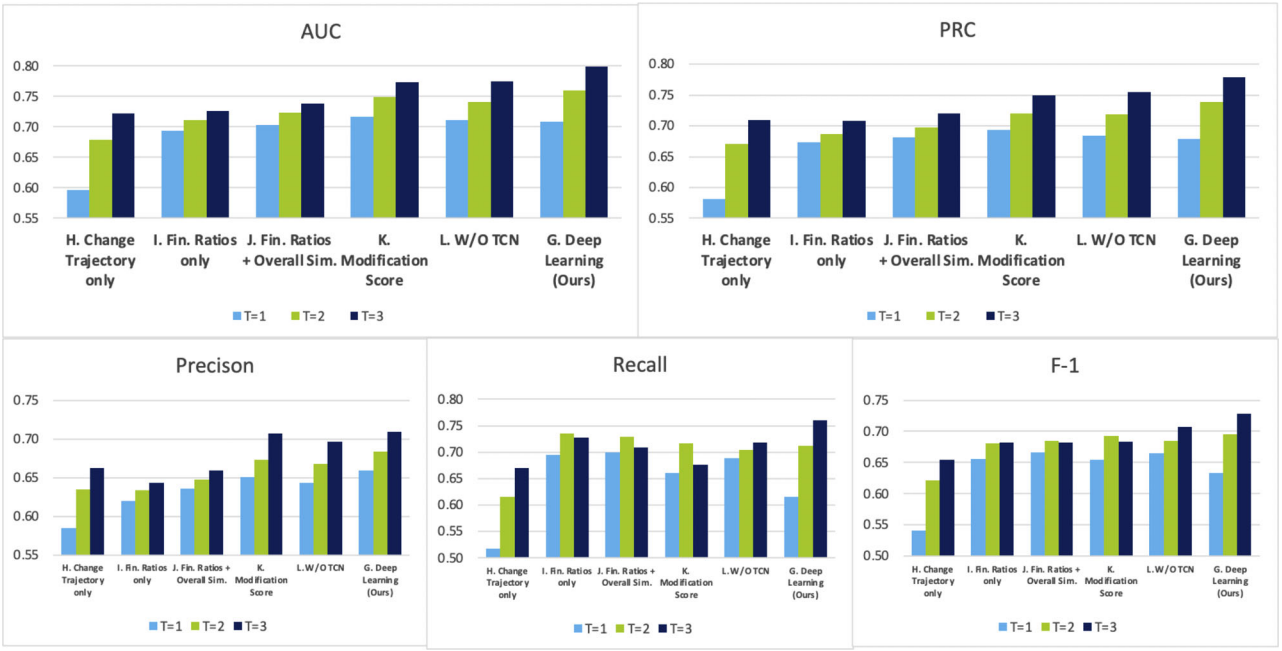

FIGURE 5 Ablation studies [Color figure can be viewed at wileyonlinelibrary.com]

change metrics, we only remove these metrics from Model E to create Model J (i.e., keep financial ratios and overall similarity as inputs). Again, AUC and PRC drop about 6% when $T = 3$. Moreover, with the overall similarity incorporated as an additional input, Model J only slightly outperforms Model I by a margin of about 1% in terms of AUC and PRC. Thus, the granular word change metrics are the major contributor to the model performance in the change trajectory. Notably, for both Models I and J, without the granular change metrics, a longer trajectory ($T = 3$) only renders a marginal gain over $T = 1$. These findings suggest that it is the word change metrics that garner useful signals in the change trajectory and boost prediction performance. With these metrics, the longer the change trajectory, the higher the model performance.

Ablation on change metrics: In this experiment, we replace the change trajectory in the deep learning model (Model G) with modification scores as proposed by Cohen et al. (2020). Recall that as modification scores are calculated without paragraph alignment, they may not properly measure changes if paragraphs have been restructured. The best results are reported in Figure 5 as Model K. Compared with Model G, the overall performance of Model K, measured by AUC and PRC, decreases by 2%–3% at $T = 3$, and by 1%–2% at $T = 2$. The Mann–Whitney U test verifies that the AUC and PRC obtained by Model G are significantly higher than those from Model I ($p < 0.01$ when $T = 3$, $p < 0.05$ when $T = 2$). This gain can be attributed to paragraph alignment that renders more accurate change measures.

Noticeably, when T decreases, the performance gain diminishes. When $T = 1$, both models G and K have similar performance as other baseline models. This is because when only two consecutive filings (i.e., $T = 1$) are compared, it can be difficult to separate truthful MD&A adjustments from manipulations. As SEC requests that MD&As must provide timely updates on business conditions, a firm always adjusts its current MD&A based on the last filing. For instance, a common MD&A topic is the comparison of revenue of the current reporting period with that of the last period. As shown in the paragraph pair 15-3 of Figure 2, such comparison must be modified accordingly every year. Therefore, changes within two consecutive filings alone may not provide sufficient evidence for frauds. In addition, the average duration of a fraud is around 3 years (Beasley et al., 2010). Thus, year-over-year change patterns over a longer time horizon can shed more light on fraudulent behaviors. For instance, under normal conditions, a firm usually makes routine adjustments in these paragraphs. Inconsistent year-over-year changes can be suspicious. Therefore, this experiment further demonstrates that our granular metrics obtained through paragraph matching are more capable of congregating subtle signals from change trajectories over a longer time horizon.

Ablation on model structure: As shown in Figure 4, it is apparent that the deep learning Model G outperforms the traditional Models A–E significantly at $T = 3$. For Models A–E, AUC remains at about 70% regardless of T, because these models are unable to encode the temporal relationships in the change trajectory. Instead, the deep learning model employs a BiLSTM layer to suitably capture the temporal patterns. Moreover, recall that our architecture has a TCN component for selecting and configuring word change metrics. In Model L, we remove TCN and feed word metrics directly to BiLSTM. The best AUC and PRC are reduced by about 3%. Finally, note that Model L is similar to Model F except a BiLSTM instead of unidirectional LSTM layer used. Model L slightly outperforms Model F by 1%–2% for all the metrics when $T = 3$. These experiments demonstrate that our

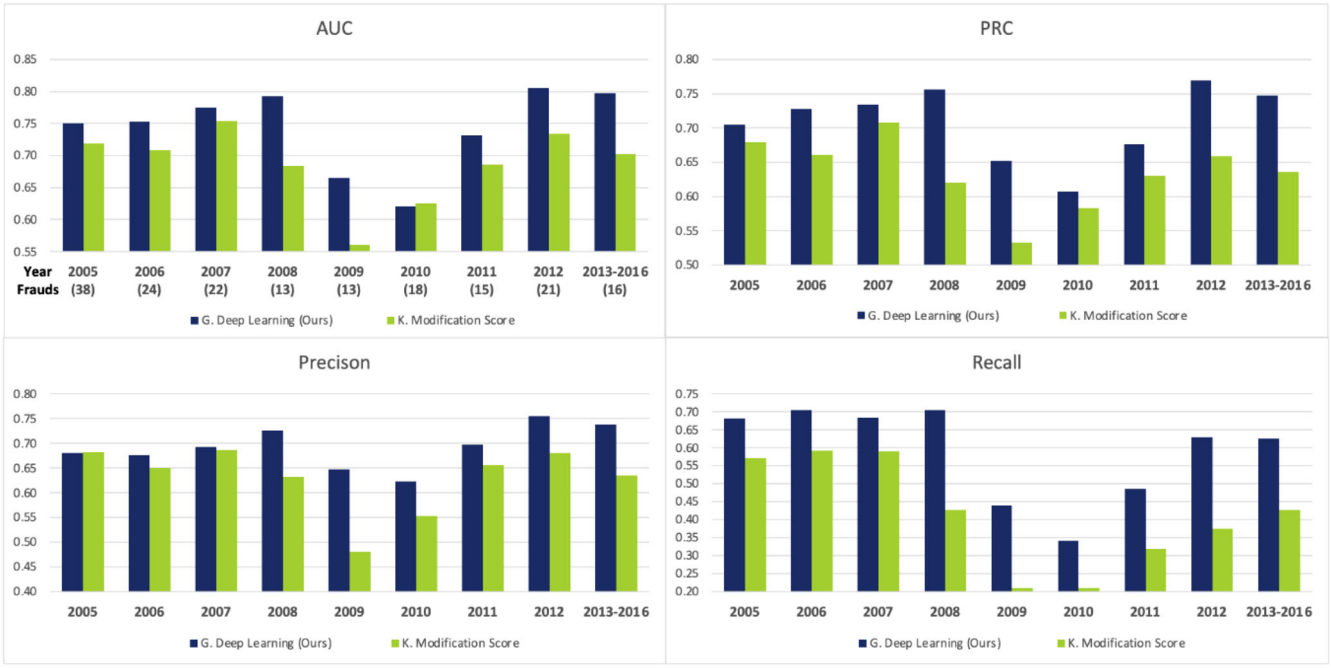

FIGURE 6 Testing model performance by walk forward validation [Color figure can be viewed at wileyonlinelibrary.com]

customized deep learning architecture can effectively extract useful signals from input variables for fraud detection.

4.2.2 | Walk forward validation

In the previous experiments, we have showcased our granular changes metrics and the deep learning model altogether can outperform all benchmark models. In this experiment, we continue to estimate the performance of our methodology in a more realistic setup. We select Model G and Model K, the two best-performing models identified by the ablation study for this test. Starting from 2005, about half of our study period, for each year t, we train these two deep learning models ($T = 3$, full inputs) using observations before year t. For example, as shown in Figure 6, all firm-year observations before 2005 are used for training and those of the year 2005 are for test. We repeat this train-test splitting process until 2013. As there have been only 16 frauds since the year 2013, we test all firm-year observations from 2013 to 2016 in one model. Note that because we train models with $T = 3$, only firms with at least three consecutive disclosures by year t are placed into the test set. Therefore, the fraud cases in Figure 6 are fewer than the numbers shown in Table 5, making this test even more challenging.

Similar to the cross-validation experiment, we train 100 models and report the average AUC, PRC, precision and recall scores on test sets. As there are very limited fraud cases in each year, in each model, we resample positive observations for 10 times and randomly sample the equal number of negative cases. For example, with 38 frauds in 2005, the test set contains 380 oversampled frauds and 380 randomly sampled negative observations. Figure 6 shows the model performance for each year under test. Overall, Model G achieves decent performance, with an average of 74% AUC, 71% PRC, 69% precision, and 59% recall,[12] while the average metrics of Model K is about 69%, 63%, 63%, and 41%, respectively. Model G outperforms Model K in almost every test set with a significant margin. In particular, for the year 2008, Model G beats Model K by 11% in AUC, 14% in PRC, 10% in precision, and 27% in recall. In contrast, the performance of Model K drops dramatically compared to the result in Figure 5. This drop may be attributed to inaccurate modification scores as the model inputs.

We can observe that both models suffer from poor performance for the years 2009–2010. This may be caused by two possible reasons. First, the number of positive observations in each test set is very small. Even one false negative can significantly affect AUC and PRC. Second, if a brand-new firm emerges in a test set with a unique change trajectory, the model has not yet acquired sufficient knowledge about the trajectory. For instance, out of 18 fraud cases in the year 2010, there are four newly established firms without data for training. As a result, the models make wrong predictions for these cases.

4.3 | Analysis of change trajectories

Our experiments show that the rich word change metrics play a critical role in predicting frauds. In particular, its impact becomes more significant when the time window T of the change trajectory is increased. Next, we apply logistic regression to understand the associations between our granular

TABLE 6 Regression of fraud risk on change metrics

	(1) Change trajectory variables only	(2) Change trajectory variables + fin. variables
const	−5.014***	−7.484***
Overall_Change $(1 - s_t)$	0.885***	0.727***
WeakModal_add (add_t^{WM})	38.469***	40.051***
WeakModal_del (del_t^{WM})	19.393***	19.480***
Negative_up (up_t^{Neg})	0.774**	1.127***
Negative_down $(down_t^{Neg})$	−0.829*	−0.100*
Litigious_del (del_t^{Litig})	9.851	−9.938
StrongModal_add (add_t^{SM})	−31.599***	−20.715*
Reward_add (add_t^{Reward})	−19.999**	−28.631***
Reward_del (del_t^{Reward})	−12.002	−14.964*
Executive_change (dummy)	No	−0.027
Fin. indicator covariates	No	Yes
Observations	87,765	87,765
Log-likelihood	−4129.2	−3940.0

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

change metrics and frauds. The dependent variable $fraud_t^k$ is 1 if the filing of firm k in year t is fraudulent, and 0 otherwise. The independent variables contain financial ratios (see Appendix 2 of the Supporting Information), overall similarity, word change metrics, and the executive change indicator. The regression can be formulated as follows:

$$fraud_t^k = \alpha + \beta\ Financial\ Ratios_t^k + \gamma\ Overall\ Similarity_t^k + \sigma\ Word\ Change_t^k + \theta\ Executive\_Change_t^k. \quad (4)$$

We only consider the change trajectory of $T = 1$ in this regression analysis because observations in a time series are typically correlated. Moreover, some word categories, for example, Weak Modal versus Uncertainty, overlap with each other. Thus, the corresponding word change metrics are also correlated. We applied forward selection during regression analysis and identified 9 variables that have significant associations with $fraud_t^k$, as shown in Table 6. We trained two regression models. Model 1 only has the selected change trajectory variables and Model 2 includes all control variables. All of the selected variables have low correlations and variance inflation factors are below 1.5, indicating multicollinearity is not an issue.

We found that the overall change $(1 - s_t)$ is positively correlated with frauds (p value < 0.01) after controlling executive changes, which actually show no significant association with frauds. Dramatic changes from the previous MD&As imply a higher risk of frauds. Changes in MD&As often convey valuable information about the movement of business activities (Brown and Tucker, 2011). Cohen et al. (2020) found that the more a firm tends to modify disclosures, the weaker financial performance it has. Thus, dramatic MD&A changes may often indicate that the firm was undergoing economic turmoil and cannot meet market expectations. As frauds are primarily driven by deteriorated performance or dramatic economic changes, managers may have incentives to manipulate information in MD&As in order to conceal performance loss or irregular behaviors (Hoberg & Lewis, 2017). Our study offers empirical evidence that MD&A changes have prediction power for frauds, in addition to their proven signaling effects on earnings, profitability, and others (Cohen et al., 2020).

Besides, we found that dishonest managers are concerned about weak modal words and are more likely to add or delete contents with low confidence and imprecise words. A higher percentage of weak modal words in newly added or deleted paragraphs indicates more risk of frauds (p value < 0.01). Interestingly, add_t^{WM} indicates increased weak modal words due to content addition, while del_t^{WM} captures the reduction of these words by content deletion. Driven by these two opposite variables, the total weak modal words in an MD&A may barely change. However, add_t^{WM} and del_t^{WM} as two granular metrics are able to capture the nuances of weak modal words used in MD&As. Previous studies have observed that deceptive statements tend to have more uncertainty and "distancing" in the language (Burgoon et al., 2003; Zhou et al., 2004). However, Humpherys et al. (2011) found that there was no difference in the use of modal words between truthful and fraudulent MD&As. Their study measures uncertainty by the ratio of modal verbs out of total verbs, which may miss subtle changes in MD&As. Our work offers a resolution to this contradiction by showing that uncertainty is strongly associated with frauds when measured by a nuanced method. Another interesting finding is that only weak modal words in the newly added or deleted contents have strong associations with frauds. This can be explained by what contents are often added or removed. Brown and Tucker (2011) indicate that new or removed contents primarily concentrate on risk factors (70%) or firm overviews (40%). The increased use of weak modal words in describing risk factors or firm overviews may indicate managers lack conviction and confidence in these statements and try to distant themselves from the statements. This lack of embracement is considered as a strong signal for deceptions (Vrij, 2008).

Moreover, in line with previous work (Goel & Uzuner, 2016; Loughran & Mcdonald, 2011), we also find that negative words are relevant to frauds. However, different from these studies that count the total negative words within an MD&A, our analysis finds that negative words in recurrent paragraphs ($up_t^{Neg}, down_t^{Neg}$) are significantly associated with frauds. Specifically, within recurrent paragraphs, an uptick in negative words implies increased risk of frauds (p value < 0.01), whereas a downtick in such words infers risk reduction (p value < 0.01). One possible explanation is that recurrent paragraphs primarily cover operations and liquidity and capital resource (LCR) (Brown & Tucker, 2011) and an increase of negative tone in the discussion may indicate deteriorating performance, which is considered as the major trigger of fraudulent

behaviors (Rezaee, 2005; SEC, 2021). By the emotion perspective, deceivers may also show increased negativity because they are afraid to be caught in a deceptive act (Vrij, 2008). In contrast, the discussion with reduced negative words may signify improvement in these routine activities. These findings offer new insights into the association between negative tone and fraudulent disclosures.

In addition, we also find reward words in either newly added or deleted paragraphs are negatively associated with frauds. The higher the values of these two variables, the lower the fraud risk. Reward words, which imply incentives driving a firm's behaviors, usually appear more frequently in contents about a firm's overview or risk factors than in operations or LCR contents, because the former can be discussed at managers' discretion, while the latter are regulated by accounting guidelines. Contents about overviews or risk factors usually have more additions or deletions (Brown & Tucker, 2011), and thus, add_t^{Reward} and del_t^{Reward}, which measure such changes, become relevant. Moreover, Jiang and Wilson (2018) found that the use of reward words is positively associated with the veracity of online posts. The authors believed these words are used to discuss concrete topics that misinformation usually lacks. These explanations can help justify our finding here. Increase reward focus implies more concrete details, which are less likely manipulated according to the cognitive effort perspective of deception theories (Vrij, 2008).

4.4 | Using alternative paragraph embedding: BERT

A critical step of our methodology is paragraph alignment based on the similarities between paragraph embeddings. We have used TF-IDF embeddings to demonstrate our methodology. Now we switch to BERT embeddings and redo all the steps in our methodology. We embed each paragraph using sentence BERT (SBERT; Reimers & Gurevych, 2019),[13] a state-of-the-art sentence embedding model. SBERT extends the pretrained BERT network using Siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity. With the paragraph embeddings, we conduct the remaining steps as described before.

As shown in Appendix 7 of the Supporting Information, we obtained comparable results as before, except that the similarity threshold is set to 0.7 (whereas 0.3 with TF-IDF). As a BERT embedding attempts to capture the overall meaning of a sentence, the cosine similarity between BERT embeddings measures the relatedness of meanings and is relatively higher than the one obtained by TF-IDF. More detailed analysis and comparison of these two embedding techniques can be found in Appendix 7 of the Supporting Information. After removing matched pairs with similarity less than 0.7, the agreement rate of the matched pairs obtained through these two embedding methods is 75%. Accordingly, the average correlation of the granular change metrics obtained through them is 0.81 (see Table A7-1 in Appendix 7 of the Supporting Information). Therefore, it is unsurprising that similar performance can be obtained in the cross-validation test. Our deep learning model still receives the highest AUC of 0.80 and PRC of 0.78 when the window $T = 3$.

5 | DISCUSSION AND CONCLUSION

Integrity and accuracy in financial disclosures are critical to the functioning of capital markets. Predictive models are much needed for stakeholders and resource-constrained SEC to flag and monitor highly probable firms. Motivated by recent studies that showed year-over-year changes of the MD&A sections contain subtle but powerful predictive signals, in this paper, we proposed a nuanced method to detect frauds by tracking granular changes in disclosures over time. We used an optimized method to align paragraphs in consecutive MD&As to locate specific changes. With paragraph alignment, we identified three types of changed contents: recurrent, newly added, and deleted contents. For each type, we measured the changes in terms of fraud-relevant linguistics features, such as sentiment, uncertainties, and award focus. Then, we represented a firm's MD&A change trajectory over years as a multivariate time series of these granular features. We developed a deep learning model to predict frauds using the change trajectory as an input. This model includes a TCN that selects and configures input metrics and a BiLSTM unit to extract predictive signals.

We conducted extensive experiments to test our method using cross-validation and walk forward validation. We benchmarked it with widely used classification models and carried out ablation studies. Our model significantly outperforms benchmark models by over 10% in terms of AUC and PRC. The ablation studies indicate that this gain can be attributed to the granular change metrics. Interestingly, this gain increases with the time span of the change trajectory. The walk forward validation showcased how our methodology can be used to detect frauds in practice. Trained by change trajectories up to year $t-1$, our deep learning model can flag high-risk cases of year t with high accuracy. Moreover, our study found specific types of changes, for example, uncertainties in newly added or deleted content, and upticks or downticks in negative words in recurrent contents, are strongly associated with frauds.

5.1 | Implications for literature

Our study offers several implications for Information Systems (IS) and OM literatures. In particular, our work falls into one of the key areas that constitute the OM and IS interface, as we apply an optimization-based matching, a technique originated from OM to solve IS problems (Kumar et al., 2018). Fraud detection has been an enduring topic in IS literature. Our study contributes to the fraud detection literature a

new predictive signal—disclosure change trajectories, along with an effective deep learning architecture. This signal is not only comprehensive as it captures rich linguistic features, but also interpretable as it allows stakeholders to trace changes in specific business activities over time. We empirically demonstrated that this model can effectively identify fraudulent cases, significantly outperforming benchmark models.

Second, our work extends the applications of the optimization-based matching technique to a new domain, text analytics. Our method can capture nuances in changes and precisely define changes trajectories. This general method can be potentially applied to other domains, such as tracing message mutation during information diffusion, discovering similarities among product variants to support product family design. In addition, recently, researchers have advocated the use of data-driven analytics to solve challenging OM issues (Choi et al., 2018). We responded to this call by conducting research on a large sample of text data and developing a reusable methodology. Moreover, as most frauds are related to the results of operations, for example, revenue, goods sold, or inventory (Dechow et al., 2011; Huang et al., 2017), accurately identifying frauds can help pinpoint issues in OM and estimate the impact of operation issues on regulatory risks. To this end, our work can also enrich research on the operations–finance interface in risk management (Wang et al., 2021).

Finally, our study contributes to the literature on misinformation. Previous studies find uncertainties and sentiment are related to misinformation (Loughran & Mcdonald, 2011; Zhou et al., 2004). Our work contextualizes these general theories in the area of financial misinformation. We found that frauds are significantly associated with uncertainties and reward focus in *newly added or deleted contents*, and the negative sentiment of *recurrent contents* has a strong indication for frauds. Previous work has given contradictory conclusions on the relationship between uncertainties and financial frauds (Humpherys et al., 2011; Zhou et al., 2004). Our findings resolve this contradiction. Overall, these observations further enhance our understanding of financial statement frauds.

5.2 | Implications for practice

SEC has committed substantial resources in policing frauds. SEC relies on tips, complaints, and referrals (TCRs) and a whistleblower program to detect wrongdoings (SEC, 2020). As a reflection of the pandemic's impact on frauds, in the year 2020, SEC received record numbers of TCRs and whistleblower tips (SEC, 2020). This placed enormous pressure on the resource-constrained regulator. The TCRs and tips need to be reviewed to identify those that warrant further investigation. On average, each investigation takes about 3 years (SEC, 2020). Despite substantial efforts in effectively triaging cases and accelerating investigation, SEC has been criticized for being an ineffective regulator, with specific concerns about its ability to identify financial reporting errors. A study investigated financial statement errors between 2005 and 2014 and found SEC was only able to catch about 50% of the errors (Kubic, 2020).

To overcome resource constraints, SEC took initiatives to develop software to examine language use in financial reports for signs of fraud (Eaglesham, 2013) and also apply risk-based data analytics (SEC, 2020). Our work resonates with these initiatives timely. We developed a solution to help financial regulators and policymakers streamline and automate the process of curbing frauds. Our framework takes an end-to-end approach, uses publicly available data without time-consuming feature engineering, and achieves superior performance. Just as we demonstrated in the walk forward validation, at each year, we can train a model using the change trajectories and financial ratios calculated from the past filings. When fed with the new filings, the model can rank the filings by the fraud risk to flag suspicious disclosures.

Moreover, different from traditional deep learning models, which are often considered as opaque black boxes, our method offers interpretability. Often, a financial statement needs to be interpreted with reference to statements in the previous periods. Content restructuring makes the side-by-side comparison nearly impossible. Our optimization-based paragraph alignment can locate paragraphs under dramatic changes, newly added paragraphs, or removed paragraphs. For example, we can construct the change trajectory for a specific topic as shown in Table 1, or trace all changes as shown in Figure 2. This allows regulators or analysts to efficiently browse the changes to understand what has happened to the firm.

5.3 | Potential new applications of our methodology

Besides fraud detection, our methodology can be used to create new applications with financial data. Cohen et al. (2020) found that MD&A changes, measured by modification scores, are strongly associated with firm profitability, stock returns, and bankruptcies. Our deep learning model can be easily adapted to new prediction tasks toward these targets. For example, we can include relevant financial indicators and word categories, recalculate change trajectories, and then change the prediction target to earnings per Share (EPS), cumulative abnormal return (CAR), or a bankruptcy indicator. In addition, we can extract change trajectories from other periodical business narratives. For instance, we can track changes in how analysts ask questions and how executives respond to repetitive or new questions in earnings calls. Druz et al. (2020) have found that executives' overall tone changes in earnings calls can predict stock returns. It would be interesting to discover these nuanced changes and evaluate how stock markets respond to them.

In the OM domain, our work may be used to exploit similarities among product and process variants based on product specifications for the purpose of reuse. Due to diverse customer needs, manufacturers are often confronted with

difficulties in dealing with frequent design changes and recurrent process variations. In align with previous work (Jiao et al., 2007; Moon et al., 2010), our method may be used to match customer needs with product functional requirements, find similarities among product variants, and cluster products based on similarities to support product family design.

Our study also provides a general approach to studying change trajectories of narratives in social media or user generated content. For example, our methods can be used to analyze how customer opinions regarding a product or product aspects change over time and this change trajectory may provide timely information that is especially helpful in prediction tasks such as defect discovery (Abrahams et al., 2015) and sale forecast (Lau et al., 2018). Another potential use is to trace the footprint of misinformation on social media. Intentional spreaders may manipulate information to make it look like real news. Shin et al. (2018) hypothesized that the need for change is particularly present for resurging rumors. This study used cosine similarity to measure the mutation of rumors. Our method can be used to capture how rumors deviate from the original version of the story over time and assess the impact of the mutation process on misinformation diffusion. In a similar vein, a study (Im et al., 2011) finds news content may also constantly evolve by adding information or changing its narrative during diffusion, while existing work on information diffusion often paid little attention to message evolution. Our work may also help foster new research in this front.

5.4 | Limitations and future opportunities

Our study has several limitations. First, although the use of AAERs as a proxy for manipulation is well-accepted in the literature (Abbasi et al., 2012; Cohen et al., 2020), AAERs only contain frauds that have been identified by SEC heretofore, and could miss frauds that have not yet been identified (or can never been identified). More frauds may be discovered in recent filings as SEC investigations are still ongoing. Nevertheless, this caveat has minimal impact on our walk forward validation, as recent MD&As were reserved for testing only. Second, we predict whether an MD&A is manipulated without considering specific fraud schemes (e.g., revenue, inventory, etc.). Future research could formulate a multi-label classification problem to predict fraud schemes. Third, in this work, we included 10 word categories identified by the previous literature. Future work could explore more features or automatically discover other relevant features. Finally, we have identified several potential applications of our methodology. Our work can be extended to future research in these areas.

ORCID

Rong Liu https://orcid.org/0000-0001-7176-1999
Zhongju Zhang https://orcid.org/0000-0001-9200-2369

ENDNOTES

[1] https://www.sec.gov/corpfin/cf-manual/topic-9

[2] The latest Accounting and Auditing Enforcement Releases (AAER) dataset compiled by Dechow et al. (2011) contains 2110 10-K frauds from 1971 to 2018. Among them, 605 frauds happened within three successive firm-years.

[3] We find that unmatched paragraphs in general have very low similarities, with an average score of 0.15. This indicates that unmatched paragraphs most likely have unique themes.

[4] The kappa score of the annotators is 0.93, indicating strong agreement.

[5] These two MD&As are taken from: https://www.sec.gov/Archives/edgar/data/1007021/0000927016-97-001813.txt, and https://www.sec.gov/Archives/edgar/data/1007021/0000927016-98-002541.txt

[6] Based on the correlations, we list the word categories in the following order: Weak Modal, Uncertainty, Negative, Litigious, Strong Modal, Reward, Compare, Positive, Discrep, Achieve.

[7] The original AAERs released by SEC record various reporting-related enforcement actions in free text (https://www.sec.gov/divisions/enforce/friactions/friactions2018.shtml). Dechow et al. (2011) manually examined AAERs to identify misstatements involving serious GAAP violations. This dataset is updated till 2018 (https://sites.google.com/usc.edu/aaerdataset/home?authuser=0).

[8] We matched COMPUSTAT data with 10-K by "fyear," "CIK," and "GVKEY."

[9] Observations were removed primarily due to unsuccessful CIK-GVKEY matching and missing financial ratios.

[10] The number of cases dropped significantly in the last 3 years. The drop may be caused by delayed AAER announcement because SEC tends to release an AAER at the end of the case history. The average delay is about 3 years (SEC 2020).

[11] In line with Cohen et al. (2020), we search in 10-Ks for words in these lists: {"CEO," "CFO," "Chief Executive Officer," "Chief Financial Officer"}, {"appoint," "elect," "new," "hire," "search"}. If a 10-K contains words from each of the lists and these words cooccur within 10 characters, we consider it mentions executive changes.

[12] Note that the recall scores are obtained at the default threshold of 0.5. This metric can be higher with a lower threshold, but the precision will decrease too.

[13] The implementation can be found at https://www.sbert.net. We use the best-performing pretrained general purpose model "all-mpnet-base-v2."

REFERENCES

Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). MetaFraud: A meta-learning framework for detecting financial fraud. *MIS Quarterly*, *36*(4), 1239–1327.

Abrahams, A. S., Fan, W., Wang, G. A., Zhang, Z., & Jiao, J. (2015). An integrated text analytic framework for product defect discovery. *Production and Operations Management*, *24*(6), 975–990.

Addawood, A., Badawy, A., Lerman, K., & Ferrara, E. (2019). Linguistic cues to deception: identifying political trolls on social media. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 15–25.

Agarwal, G. K., & Medury, Y. (2014). Internal auditor as accounting fraud buster. *IUP Journal of Accounting Research & Audit Practices*, *13*(1), 351–363.

Anti-Fraud Collaboration. (2021). Mitigating the risk of common fraud schemes: insights from SEC enforcement actions. https://www.thecaq.org/wp-content/uploads/2020/12/afc-mitigating-the-risk-of-common-fraud-schemes-2021-01.pdf

Beasley, M. S., Hermanson, D. R., Carcello, J. V., & Neal, T. L. (2010). Fraudulent financial reporting: 1998–2007: An analysis of U.S. public companies.

Beneish, M. D. (1999). The detection of earnings manipulation. *Financial Analysts Journal*, *55*(5), 24–36.

Brown, N. C., Crowley, R. M., & Elliott, W. B. (2020). What are you saying? Using topic to detect financial misreporting. *Journal of Accounting Research*, *58*(1), 237–291.

Brown, S. V., & Tucker, J. W. (2011). Large-sample evidence on firms' year-over-year MD&A modifications: Large-sample evidence on MD&A modifications. *Journal of Accounting Research*, *49*(2), 309–346.

Burgoon, J. K., Blair, J. P., Qin, T., & Nunamaker, J. F. (2003). Detecting deception through linguistic analysis. In Chen, H., Miranda, R., Zeng, D.D., Demchak, C., Schroeder, J., Madhusudan, T. (eds) *Intelligence and security informatics, lecture notes in computer science*, vol 2665 (pp. 91–101), Springer. https://doi.org/10.1007/3-540-44853-5_7

Cao, S., Jiang, W., Yang, B., & Zhang, A. L. (2020). *How to talk when a machine is listening: Corporate disclosure in the age of AI*. National Bureau of Economic Research (No. w27950).

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010a). Making words work: Using financial text as a predictor of financial events. *Decision Support Systems*, *50*(1), 164–175.

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010b). Detecting management fraud in public companies. *Management Science*, *56*(7), 1146–1160.

Choi, T.-M., Wallace, S. W., & Wang, Y. (2018). Big data analytics in operations management. *Production and Operations Management*, *27*(10), 1868–1883.

Chong, A. Y. L., Li, B., Ngai, E. W., Ch'ng, E., & Lee, F. (2016). Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach. *International Journal of Operations & Production Management*, *36*(4), 358–383.

Clarke, J., Chen, H., Du, D., & Hu, Y. J. (2021). Fake news, investor attention, and market reaction. *Information Systems Research*, *32*(1), 35–52.

Cohen, L., Malloy, C., & Nguyen, Q. (2020). Lazy prices. *The Journal of Finance*, *75*(3), 1371–1415.

Cole, C. J., & Jones, C. L. (2005). Management Discussion and Analysis: A review and implications for future research. *Journal of Accounting Literature*, *24*, 135–174.

Craja, P., Kim, A., & Lessmann, S. (2020). Deep learning for detecting financial statement fraud. *Decision Support Systems*, *139*, 113421.

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting material accounting misstatements. *Contemporary Accounting Research*, *28*, 17–82.

Dong, W., Liao, S., & Zhang, Z. (2018). Leveraging financial social media data for corporate fraud detection. *Journal of Management Information Systems*, *35*(2), 461–487.

Druz, M., Petzev, I., Wagner, A. F., & Zeckhauser, R. J. (2020). When managers change their tone, analysts and investors change their tune. *Financial Analysts Journal*, *76*(2), 47–69.

Dyer, T., Lang, M., & Stice-Lawrence, L. (2017). The evolution of 10-K textual disclosure: evidence from latent Dirichlet allocation. *Journal of Accounting and Economics*, *64*(2–3), 221–245.

Eaglesham, J. (2013, May 27). Accounting fraud targeted—With crisis-related enforcement ebbing, SEC is turning back to main street. *Wall Street Journal*, C (1).

Gerards, A. M. H. (1995). Matching. *Handbooks in Operations Research and Management Science*, *7*, 135–224.

Goel, S., & Uzuner, O. (2016). Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports. *Intelligent Systems in Accounting, Finance and Management*, *23*(3), 215–239.

Habib, A., & Hossain, M. (2013). CEO/CFO characteristics and financial reporting quality: A review. *Research in Accounting Regulation*, *25*(1), 88–100.

Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud—A comparative study of machine learning methods. *Knowledge-Based Systems*, *128*, 139–152.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Hoberg, G., & Lewis, C. (2017). Do fraudulent firms produce abnormal disclosure? *Journal of Corporate Finance*, *43*, 58–85.

Hochreiter, S., & Schmidhuber, J. (1995). Long short term memory. *Neural Computation*, *9*(8), 1735–1780.

Huang, S. Y., Lin, C.-C., Chiu, A.-A., & Yen, D. C. (2017). Fraud detection using fraud triangle risk factors. *Information Systems Frontiers*, *19*(6), 1343–1356.

Huang, S.-Y., Tsaih, R.-H., & Yu, F. (2014). Topological pattern discovery and feature extraction for fraudulent financial reporting. *Expert Systems with Applications*, *41*(9), 4360–4372.

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems*, *50*(3), 585–594.

Im, Y.-H., Kim, E., Kim, K., & Kim, Y. (2011). The emerging mediascape, same old theories? A case study of online news diffusion in Korea. *New Media & Society*, *13*(4), 605–625.

Jiang, S., & Wilson, C. (2018). Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–23.

Jiao, J., Simpson, T. W., & Siddique, Z. (2007). Product family design and platform-based product development: A state-of-the-art review. *Journal of Intelligent Manufacturing*, *18*(1), 5–29.

Jiao, J., Zhang, L., Pokharel, S., & He, Z. (2007). Identifying generic routings for product families based on text mining and tree matching. *Decision Support Systems*, *43*(3), 866–883.

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, *33*, 171–185.

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, *32*(4), 995–1003.

Ko, D., Mai, F., Shan, Z., & Zhang, D. (2019). Operational efficiency and patient-centered health care: A view from online physician reviews. *Journal of Operations Management*, *65*(4), 353–379.

Kubic, M. (2020). Examining the examiners: SEC error detection rates and human capital allocation. *The Accounting Review*, *96*(3), 313–341.

Kumar, S., Mookerjee, V., & Shubham, A. (2018). Research in operations management and information systems interface. *Production and Operations Management*, *27*(11), 1893–1905.

Larcker, D. F., & Zakolyukina, A. A. (2012). Detecting deceptive discussions in conference calls: Detecting deceptive discussions in conference calls. *Journal of Accounting Research*, *50*(2), 495–540.

Lau, R. Y. K., Zhang, W., & Xu, W. (2018). Parallel aspect-oriented sentiment analysis for sales forecasting with big data. *Production and Operations Management*, *27*(10), 1775–1794.

Lea, C., Flynn, M. D., Vidal, R., Reiter, A., & Hager, G. D. (2017). Temporal convolutional networks for action segmentation and detection. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1003–1012), Honolulu, HI.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, *45*(2–3), 221–247.

Lin, J. W., Hwang, M. I., & Becker, J. D. (2003). A fuzzy neural network for assessing the risk of fraudulent financial reporting. *Managerial Auditing Journal*, *18*(8), 657–665.

Loughran, T., & Mcdonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65.

Mahajan, A., Dey, L., & Haque, S. M. (2008). Mining financial news for major events and their impacts on the market. *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Sydney, Australia, 423–426.

Moffitt, K., & Burns, M. B. (2009). What does that mean? Investigating obfuscation and readability cues as indicators of deception in fraudulent financial reports. *AMCIS 2009 Proceedings*, San Francisco, CA, USA, 399.

Moon, S. K., Simpson, T. W., & Kumara, S. R. T. (2010). A methodology for knowledge discovery to support product family design. *Annals of Operations Research*, *174*(1), 201–218.

Moussaïd, M., Brighton, H., & Gaissmaier, W. (2015). The amplification of risk in experimental diffusion chains. *Proceedings of the National Academy of Sciences*, *112*(18), 5631–5636.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, *5*(1), 32–38.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. University of Texas at Austin.

Persons, O. S. (1995). Using financial statement data to identify factors associated with fraudulent financial reporting. *Journal of Applied Business Research (JABR)*, *11*(3), 38–46.

Purda, L., & Skillicorn, D. (2015). Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. *Contemporary Accounting Research*, *32*(3), 1193–1223.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th Inernational Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3982–3992.

Rezaee, Z. (2005). Causes, consequences, and deterrence of financial statement fraud. *Critical Perspectives on Accounting*, *16*(3), 277–298.

Ruiz, A. P., Flynn, M., Large, J., Middlehurst, M., & Bagnall, A. (2021). The great multivariate time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, *35*(2), 401–449.

Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39). Cambridge University Press.

SEC. (2003). Interpretation: commission guidance regarding management's discussion and analysis of financial condition and results of operations. Release Nos. 33-8350; 34-48960; FR-72. https://www.sec.gov/rules/interp/33-8350.htm

SEC. (2020). Sec division of enforcement publishes annual report for fiscal year 2020. https://www.sec.gov/news/press-release/2020-274

SEC. (2021). Management's discussion and analysis, selected financial data, and supplementary financial information. https://www.federalregister.gov/documents/2021/01/11/2020-26090/managements-discussion-and-analysis-selected-financial-data-and-supplementary-financial-information

Shin, J., Jian, L., Driscoll, K., & Bar, F. (2018). The diffusion of misinformation on social media: temporal pattern, message, and source. *Computers in Human Behavior*, *83*, 278–287.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.

Throckmorton, C. S., Mayew, W. J., Venkatachalam, M., & Collins, L. M. (2015). Financial fraud detection using vocal, linguistic and financial cues. *Decision Support Systems*, *74*, 78–87.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities*, John Wiley & Sons.

Wang, J., Zhao, L., & Huchzermeier, A. (2021). Operations-finance interface in risk management: Research evolution and opportunities. *Production and Operations Management*, *30*(2), 355–389.

Xiao, L. (2018). A message's persuasive features in Wikipedia's article for deletion discussions. *Proceedings of the 9th International Conference on Social Media and Society*, Copenhagen Denmark, 345–349.

Zhang, X., Du, Q., & Zhang, Z. (2022). A theory-driven machine learning system for financial disinformation detection. *Production and Operations Management*, *31*(8), 3160–3179.

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., & Nunamaker, J. F., Jr. (2004). A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, *20*(4), 139–166.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Liu, R., Huang, J., & Zhang, Z. (2023). Tracking disclosure change trajectories for financial fraud detection. *Production and Operations Management*, *32,* 584–602. https://doi.org/10.1111/poms.13888

APPENDIX A

The appendices are available in the Supporting Information that may be found online in the Supporting Information section at the end of the article.