International Journal of Accounting Information Systems 53 (2024) 100682

Contents lists available at ScienceDirect

International Journal of Accounting Information Systems

journal homepage: www.elsevier.com/locate/accinf

Accounting fraud detection using contextual language learning

Indranil Bhattacharya [*], Ana Mickovic

University of Amsterdam – Amsterdam Business School, Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Accounting fraud detection
Natural Language Processing
BERT
Information Retrieval

ABSTRACT

Accounting fraud is a widespread problem that causes significant damage in the economic market. Detection and investigation of fraudulent firms require a large amount of time, money, and effort for corporate monitors and regulators. In this study, we explore how textual contents from financial reports help in detecting accounting fraud. Pre-trained contextual language learning models, such as BERT, have significantly advanced natural language processing in recent years. We fine-tune the BERT model on Management Discussion and Analysis (MD&A) sections of annual 10-K reports from the Securities and Exchange Commission (SEC) database. Our final model outperforms the textual benchmark model and the quantitative benchmark model from the previous literature by 15% and 12%, respectively. Further, our model identifies five times more fraudulent firm-year observations than the textual benchmark by investigating the same number of firms, and three times more than the quantitative benchmark. Optimizing this investigation process, where more fraudulent observations are detected in the same size of the investigation sample, would be of great economic significance for regulators, investors, financial analysts, and auditors.

1. Introduction

Accounting fraud affects the shareholders of the fraudulent firms, as well as other participants in the capital markets. It is a widespread problem that causes significant damage in the economic market, where some estimations indicate that corporate fraud destroys 1.7% of total equity value of all U.S. publicly traded firms, which equaled to $744bn in 2020 (Dyck et al., 2021). Despite the importance of accurately identifying fraudulent companies, detecting accounting fraud in a timely manner is extremely difficult, because it requires significant effort of regulators, and this takes time and considerable financial resources. Even when such fraud has been detected, that is often after the damage has already been done, such as in well-known examples of firms as WorldCom and Enron, which finally resulted in multi-billion losses for shareholders and many employees that lost their jobs. Therefore, detecting and preventing accounting fraud is a topic of great importance to regulators, investors, financial analysts, and auditors. While extensive research has been done to detect accounting fraud using quantitative information from the financial statements (Bao et al., 2020; Cecchini et al., 2010; Dechow et al., 2011), recent studies based on textual analysis revealed that there are clues present in financial reports that can be analyzed to predict the likelihood of fraud.

However, the literature on the use of textual information from financial reports in order to detect accounting fraud is still scarce. Most of these studies focus on the investigation of the communication style used in the financial texts by capturing the tone or sentiment of the narratives (Goel et al., 2010; Purda and Skillicorn, 2015). Recent studies also started to investigate the underlying

[*] Corresponding author.
E-mail addresses: i.bhattacharya@uva.nl (I. Bhattacharya), a.mickovic@uva.nl (A. Mickovic).

https://doi.org/10.1016/j.accinf.2024.100682
Received 19 July 2022; Received in revised form 5 September 2023; Accepted 1 March 2024
Available online 11 March 2024

topics mentioned in the texts that would indicate the possibility of misreporting (Brown et al., 2020; Minhas and Hussain, 2016).

While prior literature helps understand the facets of fraudulent texts, contemporary research argues that commonly used linguistic measures cannot adequately capture the context of management disclosures (Bushee et al., 2018; Loughran and McDonald, 2011; Loughran and McDonald, 2016), thereby limiting the inferences drawn from these measures. For example, when using *bag of words* methods such as LDA (Latent Dirichlet Allocation) for text analysis, the order of words in the text is not taken into account, therefore making the performance of such methods invariant to word permutations within each document. In contrast to this, the contextual analysis also encompasses the information on surrounding conditions and environment which are improving the understanding of the context surrounding the text. Moreover, Loughran and McDonald (2016) and Pratt (2015) called for research on the possibility of using deep learning based methods, where machine learns from enormous cloud-based data sets in order to capture the deeper meaning of the business text. However, the empirical research is still scarce. Therefore, we investigate whether a machine learning model that learns from the contexts present in the financial reports improves accounting fraud detection beyond what can be achieved by existing textual and quantitative models. Specifically, we address the following research questions:

RQ1: Does contextual learning from financial reports improve accounting fraud detection, relative to extant textual methods?

RQ2: How does contextual information supplement information obtained from existing quantitative methods?

We use Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) in order to detect fraudulent firms. BERT is a neural network model that is designed to learn the context of a language using textual inputs. It has been extensively employed in language-specific studies, including automation of question answering (Yang et al., 2019; Alberti et al., 2019) and language translation (Yang et al., 2020; Zhu et al., 2020). Recently, Liu et al. (2022) explored the relationship between Key Audit Matter sentiment and firm performance and found positive correlations, supporting the use of BERT model for sentiment extraction. In our study, we implement the BERT model in order to learn and capture the underlying contexts in the texts of financial reports. Finally, we train it to classify fraudulent firms' disclosures.

Our data is based on texts from the Management Discussion and Analysis (MD&A) section of annual 10-K reports from the Securities and Exchange Commission's (SEC) database. This section is commonly used by investors, and is recognized in the literature as an instrument that signals financial distress to investors (Holder-Webb and Cohen, 2007). To address our research questions, we use two models from the previous literature as benchmark models, namely textual and quantitative benchmark model. We fine-tune the BERT model and also construct the ensemble model based on quantitative data from financial statements, along with textual data. We show that our final model outperforms the textual benchmark model and the quantitative benchmark model from the previous literature by 15% and 12%, respectively.

Since it is, because of time and financial constraints, unrealistic for regulators and corporate monitors to investigate all publicly traded firms for accounting fraud, we measure how many fraudulent 10-K filings are being captured in the top 1% predicted observations with the highest likelihood of being fraudulent. Finally, our model identifies five times more fraudulent observations than the textual benchmark and three times more than the quantitative benchmark upon investigating the same number of firms. We, therefore, conclude that our model has a higher economic significance than the models used in the previous literature. We also perform a battery of robustness tests, to increase the confidence in our findings.

In summary, our main contributions are the following. First, we apply and fine-tune the BERT model to the accounting field and employ it in order to detect accounting fraud in publicly traded U.S. firms. Second, we show that context in the financial texts contains important information that helps detect accounting fraud. Our final model outperforms the textual benchmark model and the quantitative benchmark model from the extant literature by 15% and 12%, respectively. Third, we detect how many fraudulent filings are in the top 1%, and show that our model identifies five times more fraudulent observations than the textual benchmark by investigating the same number of firms, and three times more than the quantitative benchmark. Fourth, by uncovering specific features that make it easier or harder for BERT to detect fraud, we present practical insights for financial investigators that could support them in their decision when to pursue the investigation.

The rest of the paper is organized as follows. In Section 2, we first introduce the related work. Then, we discuss our data construction and present the experimental setup in Sections 3 and 4. In Section 5, we present our results, and in Section 6 we provide practical insights for financial investigators. In Section 7, we show supplementary analysis, and finally, we discuss future research and conclude the paper in Section 8.

2. Related work

Over the last decade, researchers explored the predictive potential beyond the quantitative information of financial statements to predict financial anomalies like misreporting and bankruptcy. The main premise of this literature is to find patterns in the textual communication to describe managers' deliberate attempts to manipulate reporting. Rogers et al. (2011) examined the disclosure tone and shareholder litigation using firms' earnings announcement. Larcker and Zakolyukina (2012) developed a linguistic model to detect deceptions in earnings conference calls. We review studies that focus on exploring the textual contents of companies' financial reports to investigate financial irregularities. Readers can also consult Loughran and McDonald (2016) for an extensive literature review that covers advances in the accounting field until the year 2016.

One stream of research examines the ease of reading financial texts. Li (2008) implemented the FOG index to measure the readability score of annual reports and found that firms with higher readability scores have more persistent positive earnings. Goel and Gangolly (2012) argued that the likelihood of fraud increases with the increasing complexity of sentences present in the financial reports. Goel et al. (2010) found that fraudulent annual reports contain more passive-voice sentences and are more difficult to read than the non-fraudulent annual reports. Similarly, Humpherys et al. (2011) investigated linguistic cues from financial disclosures, discovering that fraudulent disclosures use various cues such as activation language, imagery, and words, but less lexical diversity than the non-fraudulent statements.

Further, studies focus on deriving numerical features from the textual contents of financial reports and use them in a classifier to train a fraud detection model. Goel et al. (2010) extracted features from annual 10-K filings using the bag-of-words approach and implemented an SVM (Support Vector Machine) model to detect fraudulent activities. They also developed a list of detrimental words in the financial texts that help in separating fraud from non-fraud filings. Cecchini et al. (2010) used TF-IDF features from the MD&A section of annual 10-K reports developing an SVM model for predicting financial frauds and bankruptcy events. Glancy and Yadav (2011) proposed a quantitative fraud detection model using singular value decomposition on MD&A text data. Purda and Skillicorn (2015) studied the temporal change of annual and quarterly financial narratives using 200 most predictive words as features in an SVM model. Minhas and Hussain (2016) compared several classification algorithms after extracting n-gram features from narrative sections of annual 10-K reports. They also compared text readability tools for potential feature extraction from the documents.

Some studies particularly focus on finding patterns in topics and word combinations in order to investigate abnormal behavior. Moffit et al. (2010) derived the lexical bundles that are most frequently present in the management discussion and analysis section of the annual 10-K reports. Loughran and McDonald (2011) created a new financial dictionary and found that negative language in financial reports is associated with accounting misconduct. They developed a list of negative words that can be used to comprehend the tone and the sentiment of the annual 10-K reports. Hoberg and Lewis (2017) deployed topic modeling using LDA to find that fraudulent managers use abnormal verbal tone while writing financial disclosures. Brown et al. (2020) studied the incremental contributions of thematic contents of financial narratives using topic modeling. Both Hoberg and Lewis (2017) and Brown et al. (2020) composed extensive lists of topics that help in detecting accounting frauds. A recent study by Berkin et al. (2023) investigates the use of machine learning methods in analyzing attributional content and framing in corporate reporting. They test five machine learning classifiers on a dataset of management commentary reports to identify performance related statements, detect attributions, and classify the content of these attributions on both intra- and inter-sentential level. The results demonstrate the potential of machine learning in streamlining narrative disclosure analysis by providing an efficient method for detecting and classifying performance related attributional statements.

Recently, Craja et al. (2020) proposed a deep learning based approach to detect accounting frauds. Their study uses Hierarchical Attention Network (HAN) which utilizes structured hierarchy of MD&A sections. The model also allows for the use of attention mechanisms on both word and sentence levels, thereby providing indicators that could help stakeholders identify whether further investigation is needed. We build on the previous literature, and use a transformers based BERT model which is designed to capture contextual aspects from the text. Additionally, our study aims to provide pragmatic contributions by evaluating the model by practical measures such as NDCG@k and comparing the economic significance of the predictions.

Despite the current improvements in the natural language processing that could help understand the underlying communication style in financial statement texts, the majority of previous papers leverage either word categorizations (or dictionaries) or the discrete topics in the texts. There have been few studies concerning understanding the deeper meaning of the narratives and preserving the context of the writing. Our aim is to contribute to the existing literature by applying a model that could capture the contexts in the financial reports, and utilize this model in accounting fraud detection.

3. Data and sampling design

Our experiment is based on texts from annual 10-K reports issued by U.S. firms between years 1994 and 2013. We retrieve texts from Item 7, namely the Management Discussion and Analysis (MD&A) section of annual 10-K reports from the Securities and Exchange Commission's (SEC) EDGAR database and parse the information from Item 7 into a machine-readable format. We use MD&A section to extract texts because analyzing the content of this section is a common practice for investors seeking informational advantage (Durnev and Mangen, 2020; Loughran and McDonald, 2016; Muslu et al., 2015). Holder-Webb and Cohen (2007) indicate that the MD&A section of 10-K reports is officially recognized as a source that contains valuable information which signals financial distress to investors. Additionally, since the contents of the MD&A section is unregulated and unstructured, but highly informative about the firm (Feldman et al., 2010), we develop a contextual machine learning model that analyzes information content in the MD&A section.

The fraud indicators used in this paper are derived from the detected material accounting misstatements disclosed in the SEC's Accounting and Auditing Enforcement Releases (AAERs) provided by the USC Marshall School of Business (previously Berkeley Center for Financial Reporting and Management – CFRM) (Dechow et al., 2011). Recent literature identifies this as a leading database that contains a comprehensive list of accounting fraud cases (Karpoff et al., 2017). Other terms such as earnings management, manipulation, and misstatements, are often used interchangeably, even though the SEC often implies fraud in their allegations. Some important misstatement indicators identified by the SEC include: misstated revenue, misstatement of other expense/shareholder equity account, capitalized costs as assets, misstated accounts receivable, misstated inventory, misstated cost of goods sold, misstated reserve account, misstated liabilities, misstated marketable securities, misstated allowance for bad debt, misstated payables.

Our accounting fraud detection study is based on publicly traded U.S. firms. We construct two different datasets to address RQ1 and RQ2. To address RQ1, we construct the first dataset (referred to as the text data) based on raw texts from Item 7 of Management's Discussion and Analysis (MD&A) section of annual 10-K reports collected from the SEC EDGAR database. To address RQ2, we construct the second dataset (referred to as the ensemble data) based on features extracted from the Compustat data from the year 1994 to 2013, in order to obtain the quantitative features. We further combine these quantitative features with the text data mentioned above, and finally obtain ensemble data.

Our final datasets span from the year 1994 until the year 2013. We use 1994 as the starting year since 10-K filings are available from that year on the SEC website. Despite the USC Marshall School of Business dataset containing all AAERs issued pertaining to misreporting before the end of 2018, we have chosen to utilize 2013 as the cutoff date. This is due to the fact that it takes several years for the SEC to thoroughly investigate presumed cases of accounting fraud (Karpoff et al., 2017). Furthermore, the dataset also reveals an absence of misreporting cases during 2017 and 2018, which further emphasizes the prolonged investigation period required for such cases. Specifically, Dyck et al. (2010) find that the average time gap between the misreporting and initial detection of accounting fraud is two years. Hence, we find it necessary to limit our data until 2013 in order to ensure the reliability of the data.[1] This choice also ensures comparability with prior research. For example, Brown et al. (2020) focused on misreported 10-K reports until 2010, while Bao et al. (2020) utilized a sample until 2008. These studies also emphasize the risk of including data from subsequent years, which could inaccurately classify fraudulent 10-K reports as non-fraudulent. In the next two subsections, we discuss the construction procedure of text data and ensemble data.

3.1. Text data construction

The process of collecting the text data is presented in Fig. 1. First, from the SEC website we collect the list of all CIKs (Central Index Keys), which are unique for each publicly traded U.S. firm. For each CIK, we collect the filing dates of annual 10-K reports for 20 years in total, spanning from the year 1994 until 2013, along with the corresponding accession numbers (an_i in the Fig. 1), which are unique for each 10-K report. For each 10-K filing, we create the URL using CIK and the accession number that leads to the corresponding 10-K report. Following the method developed by Berns et al. (2022), the text parsing algorithm searches for the term “Item 7. Management Discussion and Analysis”, and any one of the phrases “the following discussion”, “this discussion and analysis”, “should be read in conjunction”, “should be read together with”, “the following management’s discussion and analysis” in the following five sentences, in order to identify the beginning of the MD&A section of 10-K reports. The end of the MD&A section is determined by searching the variations of “Item 8. Consolidated Financial Statements”.

Next, we find the list of fraudulent firm-year observations using the AAER data and link it to their corresponding CIKs. We find altogether 289 fraudulent firm-year observations,[2] and create a binary fraud flag ($fraud = 1$ for fraudulent 10-K filing, otherwise 0) as an input for our classification algorithms. Our final text data contains 30,876 firm-year observations with 289 fraudulent observations spanning between years 1994 and 2013. We find that the average MD&A section contains 8,619 words, 617 sentences, and 16 words per sentence. Panel A of Table 1 represents the yearly distribution of fraudulent observations in the text data.

3.2. Ensemble data construction

Our ensemble dataset is based on quantitative data from the Compustat database, along with text data described in the previous section. Following earlier literature, we use a list of 28 raw financial features, as adopted in the previous research (Bao et al., 2020). Using readily available information from financial statements helps with the simplification of the fraud detection process, since it avoids calculations of more complex accounting ratios. The 28 financial features can be divided into four groups based on the source of information. Those are the items that originate from balance sheets, income statements, cash flow statements, and market value items.[3]

To implement the quantitative model, we collect 28 raw financial features mentioned above, along with their corresponding CIKs, between the years 1994 and 2013. In order to test our second research question, we merge the Compustat data with the text data from Section 3.1 to obtain the ensemble data. This means that the firm-year observations in ensemble data are essentially a subset of the observations in text data.[4] Our final ensemble data contains 25,853 firm-year observations with 283 fraudulent observations. Panel B of Table 1 presents the yearly distribution of fraudulent filings in the ensemble data.

[1] The number of fraudulent observations for text data for the years 2014, 2015, 2016, 2017, and 2018 were 6, 2, 1, 0, and 0 respectively. Including those years would potentially introduce noise into our model and weaken its overall effectiveness.

[2] The parsing algorithm captures MD&A text for 219 fraudulent observations. We encounter some anomalous 10-K reports where the parsing algorithm does not work, for example, Item 8 is not found in the 10-K reports, the MD&A section is wrongly listed under Item 6, etc. In order to maximize the number of fraud observations in our data, we manually capture the remaining 70 MD&A sections from 10-K reports of fraudulent filings.

[3] The information from balance sheets include 17 variables: Current assets, Property, plant, and equipment, Accounts payable, Cash and short-term investment, Related earnings, Inventories, Common/ordinary equity, Debt in current liabilities, Receivables, Assets, Long-term debt, Current liabilities, Income taxes payable, Investment and advances, Liabilities, Short-term investments, Preferred/preference stock (capital), from income statement 7 variables: Cost of goods sold, Income before extraordinary items, Depreciation and amortization, Interest and related expense, Income taxes, Sales/turnover (net), Net income (loss), from cash flow statement 2 variables: Sale of common and preferred stock, Long-term debt issuance, and from market-value 2 items: Common shares outstanding, Price close.

[4] We merge the Compustat data and text data using CIK and year as merging keys. As a result, the number of fraudulent observations in the ensemble data decreases by 6 compared to the text data, due to the missing data in the merging process.

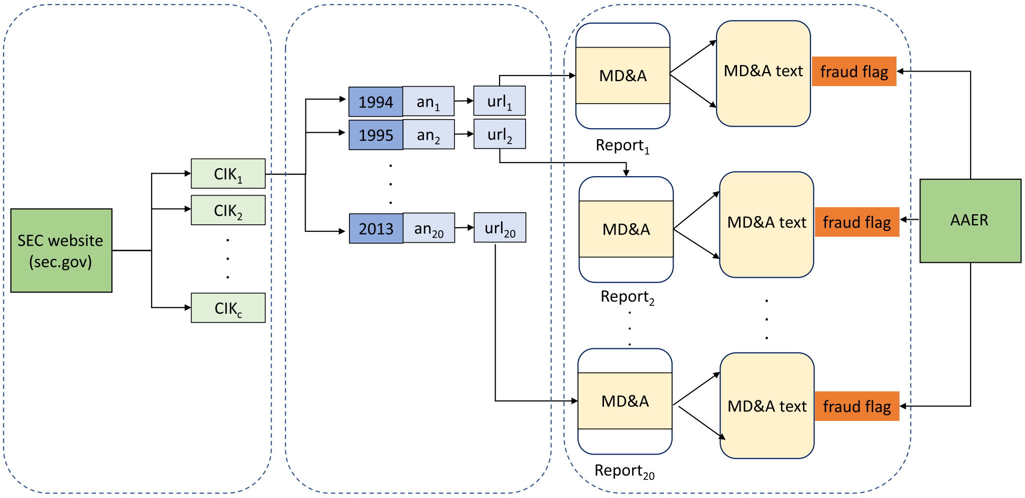


Fig. 1. Data collection process.

Table 1
Yearly distribution of fraudulent 10-K filings.

	Panel A: Yearly distribution of fraudulent 10-K filings in text data			Panel B: Yearly distribution of fraudulent 10-K filings in ensemble data		
Year	Total number of filings	Number of fraud filings	Percentage of fraud filings	Total number of filings	Number of fraud filings	Percentage of fraud filings
1994	161	1	0.62%	147	1	0.68%
1995	211	2	0.95%	194	2	1.03%
1996	343	4	1.17%	301	4	1.33%
1997	542	10	1.85%	486	10	2.06%
1998	626	13	2.08%	552	13	2.36%
1999	729	14	1.92%	642	14	2.18%
2000	794	22	2.77%	706	22	3.12%
2001	859	25	2.91%	768	25	3.26%
2002	1,012	31	3.06%	905	31	3.43%
2003	1,438	32	2.23%	1,282	32	2.50%
2004	1,523	24	1.58%	1,374	24	1.75%
2005	1,625	20	1.23%	1,468	20	1.36%
2006	1,775	13	0.73%	1,592	13	0.82%
2007	1,883	10	0.53%	1,685	9	0.53%
2008	2,167	8	0.37%	1,798	7	0.39%
2009	2,767	11	0.40%	1,905	9	0.47%
2010	2,840	11	0.39%	1,952	11	0.56%
2011	3,049	12	0.39%	2,084	12	0.58%
2012	3,194	16	0.50%	2,161	16	0.74%
2013	3,338	10	0.30%	2,282	8	0.35%

4. Experimental setup

4.1. Method

We use the BERT-Base model (uncased)[5] from TensorFlow Hub (Abadi et al., 2016) which has been pre-trained for the English language using Wikipedia and BookCorpus. Furthermore, we use WordPiece tokenizer that creates tokens, elementary lexical components, by splitting the text into words on punctuation and white spaces, and further tokenizing words into word pieces.

Following Devlin et al. (2019), we also use a special classification token [CLS] in the beginning and a separation token [SEP] at the end of every input text sample sequence. BERT-Base model uses 12 hidden layers of transformer blocks with hidden dimension of 768

[5] https://tfhub.dev/tensorflow/bert_en_uncased_L-12_H-768_A-12/4.

and 12 attention heads. For each BERT-Base model, we use the maximum sequence length of 512 tokens of texts including the [CLS] and [SEP] tokens. We add an output sigmoid layer at the end of last layer of the BERT-Base model in order to establish the rank of predictions indicating the likelihood of fraud. Our search space to find optimal hyperparameters are also based on Devlin et al. (2019)'s fine-tuning strategy. While we use a fixed learning rate of 1e-5 and adam optimizer, our optimal batch size and number of epochs are found from searching within the set {1,2,..,8}.

We use the rank average of the model predictions because a fraud prediction calculation can be viewed as a ranking task. It is important to understand whether the fraudulent observations are ranked higher in the probability of being fraudulent than the non-fraudulent observations. Our chosen evaluation metrics are also aligned with this choice (see Section 4.3). Additionally, rank averaging can be performed regardless of the scales of predictions while reducing the variance of the final prediction. Unlike simple averaging, rank averaging is scale-independent and reduces the variance of the final prediction.

To illustrate the rank average method, let us consider a hypothetical example with six observations with true fraud flags of [1, 0, 1, 0, 1, 0], where 1 represents fraudulent observations and 0 represents non-fraudulent observations. Assume that two predictions are obtained from two models: [0.8, 0.3, 0.6, 0.5, 0.4, 0.35] and [0.6, 0.9, 0.3, 0.4, 0.95, 0.2], where the numbers in vectors indicate the probability of an observation to be fraudulent. The rank vectors of these two predictions and their ranked observations based on the possibility of being fraudulent are [6, 1, 5, 4, 3, 2] and [4, 5, 2, 3, 6, 1], respectively. The individual AUC scores of these two models are 0.889 and 0.667, respectively. The simple average of these two models' predictions would be [0.7, 0.6, 0.45, 0.45, 0.68, 0.28], which obtains an AUC score of 0.778. However, the rank average of these two predictions would involve calculating the average for each observation from the rank vectors and would result in [5, 3, 3.5, 3.5, 4.5, 1.5], which obtains an AUC score of 0.945.

Our final model $BERT_{final}$ is the rank average of predictions from two separate fine-tuned BERT models: $BERT_{first}$ and $BERT_{last}$, where $BERT_{first}$ is trained on the first 512 tokens of each text samples and $BERT_{last}$ is trained on the last 512 tokens of each text sample. The process of training this model from two sources is illustrated in Fig. 2. The choice of including first and last tokens from each document is further supported by Sun et al. (2019), who show that including both beginning and the end of an article results in lower error rates. Whereas initial tokens in the MD&A reports usually contain introductory comments, the final tokens contain concluding remarks of the MD&A section and mostly provide company's vision for the future and next steps. Since those two sources provide fundamentally different information, both of which are equally important, we ensemble predictions of these two models.

The introductory notes of financial reports usually provide a concise summary of a company's financial performance in the previous year, including a comprehensive overview of financial conditions and Results of Operations. However, some companies may resort to generic template sentences, which goes against the guidelines set forth by the SEC and happens less frequently. According to the SEC, the introduction of the MD&A section should provide useful information and avoid boilerplate language that merely duplicates the more detailed analysis that follows. As they state, "boilerplate disclaimers and other generic language generally are not helpful in providing useful information or achieving balance, and would detract from the purpose of the introduction or overview" (Securities Exchange Commission, 2003).

The Online appendix includes four examples of MD&As and their first and last 512 tokens. Although some of the examples use generic text, it can be difficult to discern where it starts and stops. While some of the information about the company's structure may be generic, it could also be vital for stakeholders. The first 512 tokens contain an average of 16 words per sentence, translating to around 32 sentences, which usually comprise 4–5 paragraphs. While it may be tempting to remove some sentences, doing so could lead to the loss of important information, and the "noisy" information could be valuable to the model. Therefore, it is more beneficial to include all available information, even if it may not seem immediately relevant, than to risk losing potentially crucial input. In addition, we have observed a significant increase in the number of words used in MD&A sections over the years. For example, in 1994, the average number of words used in MD&A was around 3,500, whereas by 2013, it had increased to over 9,000. To ensure comparability between reports, it is essential to consistently examine the information contained in both the introductory and conclusion notes.

Our observations are consistent with previous literature, such as the study by Siano and Wysocki (2021), which used only the first 10 sentences from each earnings announcement due to limitations in computing resources. Additionally, Sun et al. (2019) found that the most important information is typically summarized in the beginning or end of the text, albeit on the IMDb dataset. Based on this literature, we hypothesize that the critical information in MD&A sections is likely located in these sections and therefore we focus our analysis on these sections.

For prediction, we use rank average of predictions from $BERT_{first}$ and $BERT_{last}$ as described in Fig. 2. For each of the 10-K reports in the set $\{Report_1, Report_2, \ldots, Report_n\}$, we first extract MD&A section. Then we use first 512 and last 512 tokens of MD&A sections as inputs to train $BERT_{first}$ and $BERT_{last}$ respectively. We obtain the $pred_{first}$ as output prediction from the sigmoid output layer of $BERT_{first}$ and obtain $pred_{last}$ from $BERT_{last}$. Our final prediction of $BERT_{final}$ is finally the rank average of $pred_{first}$ and $pred_{last}$, that is $pred_{final} = 0.5 * rank(pred_{first}) + 0.5 * rank(pred_{last})$.

4.2. Validation strategy

We use rolling windows of consecutive 5 years to train our models, and the immediate following year as the test set, in order to evaluate the performance of the models. We use the period between the year 1994 and 1999 as our validation set. Specifically, to optimize parameters in the models, we initially train our models on the years between 1994 and 1998, to predict on the year 1999. Each model is trained with a different combination of parameters. After that, with the final set of parameters that produces the optimum prediction on 1999, we train our models on every 5-years data and predict accounting fraud on the immediate next year. This procedure is presented in Fig. 3. From the Figure, it is visible that our first training period is the interval between the years 1995 and

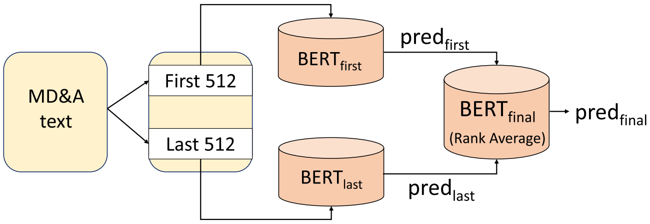


Fig. 2. Model training from input text.

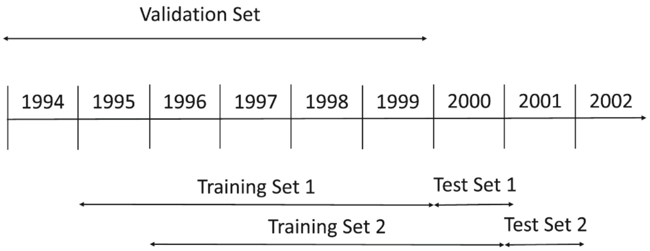


Fig. 3. Validation strategy of our model.

1999, and the corresponding test set is for the year 2000. Similarly, our second training period is the interval between 1996 and 2000, and the corresponding test set is the year 2001. This procedure continues until the end of our sample, which results in 14 years of test period between years 2000 and 2013. Importantly, this strategy is also in line with the previous research Brown et al. (2020), which allows us to compare our model against the benchmarks established in the literature.

4.3. Evaluation metrics

We use the area under the ROC curve (AUC) as our primary evaluation metric. Fraud detection models suffer from class imbalance problems, which makes AUC a reasonable metric choice, as it presents the probability that a randomly selected fraud sample would be ranked higher than a randomly selected non-fraud sample.

As our second evaluation metric, we use normalized discounted cumulative gain at the position k (NDCG@k). Since the task of fraud prediction can also be postulated as a ranking problem, the NDCG@k provides insight into the structure of top k observations that have the highest probability of being fraudulent and that is in agreement with the original fraud samples. NDCG@k represents the ratio where a higher value represents better performance, and the measure ranges from 0 to 1. While the NDCG@k value of 1 represents that the first k observations with the highest prediction scores of being fraudulent are all true fraud samples, the NDCG@k value of 0 would indicate that none of the first k observations with the highest prediction scores of being fraudulent are true fraud samples. Throughout our study, we use 1% of firms in each test year to report the NDCG@k scores.

Because of time and financial constraints, it is unrealistic for regulators and corporate monitors to investigate all publicly traded firms for accounting fraud, we also measure in absolute terms how many fraud samples are being captured in the top 1% predicted firms (highest likelihood of being fraudulent). This metric, along with the NDCG@k measure, helps us evaluate the economic significance of the models, identifying if more fraudulent firms could be captured by investigating the same number of firms. We refer to this measure as *Capture* in Tables 2 and 3, where we present the performance results of our models.

Recent accounting fraud detection research (Bao et al., 2020; Brown et al., 2020) uses AUC and NDCG@k in their studies. Hence, using those metrics as reference points provides a valid comparison of our models against the benchmarks.

5. Results

To address our research questions and to provide support for our analysis, we use two models from the previous literature as benchmark models. We, therefore, compare the performance of our final models against the existing benchmark models. We use the Latent Dirichlet Allocation (LDA) model as our textual benchmark model and the RUSBoost model as our quantitative benchmark model for accounting fraud prediction. We use LDA as our textual benchmark model since recent studies by Brown et al. (2020) and Hoberg and Lewis (2017) demonstrate that the LDA model outperforms the commonly used approach using textual style features to detect accounting frauds. On the other hand, we use the RUSBoost model as our quantitative benchmark model, since Bao et al. (2020) show that the RUSBoost model outperforms commonly used logistic regression used by Dechow et al. (2011) and support vector machine from Cecchini et al. (2010) to detect accounting frauds.

Table 2
Yearly performance of LDA and BERT models on the textual data. AUC – area under the ROC curve, LDA – Latent Dirichlet Allocation, NDCG@k – normalized discounted cumulative gain at the position k, BERT – Bidirectional Encoder Representations from Transformers, Average – average across 14 test years.

AUC	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	Average
LDA	0.793	0.708	0.684	0.697	0.740	0.715	0.726	0.708	0.749	0.684	0.767	0.803	0.711	0.600	0.720
$BERT_{first}$	0.833	**0.810**	0.804	0.812	0.867	**0.887**	0.760	0.814	0.791	**0.736**	0.799	0.803	0.749	**0.787**	0.804
$BERT_{last}$	0.831	0.759	**0.879**	**0.896**	0.848	0.800	**0.849**	0.868	**0.858**	0.664	0.780	**0.881**	0.751	0.757	0.816
$BERT_{final}$	**0.845**	0.809	0.865	0.876	**0.871**	0.858	0.824	**0.874**	0.842	0.682	**0.818**	0.864	**0.769**	0.774	**0.826**

NDCG@k	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	Average
LDA	0.000	0.747	0.000	0.000	0.139	0.071	0.126	0.197	0.165	0.054	0.053	0.196	0.034	0.000	0.127
$BERT_{first}$	**1.000**	**1.000**	**1.000**	**1.000**	0.843	0.591	0.492	**0.616**	**0.456**	**0.539**	0.219	0.546	0.505	**0.432**	0.660
$BERT_{last}$	0.920	**1.000**	**1.000**	**1.000**	0.751	0.550	0.413	0.587	0.326	0.500	0.393	0.477	0.415	0.411	0.624
$BERT_{final}$	**1.000**	**1.000**	**1.000**	**1.000**	**0.852**	**0.637**	**0.504**	0.522	0.398	**0.539**	**0.482**	**0.592**	**0.522**	0.403	**0.675**

Capture	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	Sum
LDA	0	6	0	1	2	1	2	2	2	1	1	1	1	0	20
$BERT_{first}$	**8**	**9**	**11**	14	12	8	5	**5**	3	**6**	3	5	**6**	**5**	100
$BERT_{last}$	7	**9**	**11**	**15**	11	7	5	**5**	2	5	4	6	5	**5**	97
$BERT_{final}$	**8**	**9**	**11**	**15**	**13**	**10**	**6**	4	**4**	5	**5**	**7**	**6**	4	**107**

Table 3

Yearly performance of RUSBoost, BERT and Ensemble model on the ensemble data. AUC – area under the ROC curve, NDCG@k – normalized discounted cumulative gain at the position k, BERT – Bidirectional Encoder Representations from Transformers, Ensemble – ensemble model based on both quantitative and textual data, Capture – the number of fraudulent observations that could be captured by investigating the same number of observations, Average – average across 14 test years.

AUC	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	Average
RUSBoost	0.784	0.780	0.808	0.830	0.823	0.802	0.671	0.683	0.768	0.735	0.750	0.865	0.723	0.620	0.760
$BERT_{final}$	0.840	0.804	0.863	0.874	0.868	0.853	**0.820**	**0.947**	**0.900**	**0.814**	0.822	0.867	0.765	**0.897**	**0.852**
Ensemble	**0.855**	**0.821**	**0.880**	**0.903**	**0.878**	**0.867**	0.781	0.882	0.857	0.813	**0.824**	**0.903**	**0.769**	0.827	0.847

NDCG@k	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	Average
RUSBoost	0.586	0.731	0.673	0.473	0.000	0.043	0.000	0.000	0.373	0.176	0.265	0.274	0.316	0.000	0.279
$BERT_{final}$	**1.000**	**1.000**	**1.000**	**1.000**	**0.846**	**0.663**	**0.512**	**0.575**	0.371	**0.611**	0.491	0.511	**0.522**	**0.468**	**0.684**
Ensemble	**1.000**	**1.000**	0.929	0.660	0.594	0.343	0.207	0.239	**0.586**	0.383	**0.557**	**0.607**	0.498	0.090	0.550

Capture	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	Sum
RUSBoost	3	6	5	5	0	1	0	0	2	2	2	3	3	0	32
$BERT_{final}$	**8**	**8**	**10**	**13**	**10**	**9**	**6**	**4**	**3**	**5**	**5**	5	**6**	**4**	**96**
Ensemble	7	**8**	9	7	9	5	3	**4**	**3**	2	**5**	**6**	**6**	1	75

5.1. Addressing RQ1

To examine whether contextual learning from financial reports improves accounting fraud detection relative to the extant textual method, we compare our BERT models and the LDA benchmark model using the text data. The results of the calculations are presented in the Table 2.

First, we discuss the performance of the LDA model. Similar to Brown et al. (2020) and Hoberg and Lewis (2017), we adopted disintegrated topic features as vectors to use them in a logistic regression model as shown in Eq. 1. We used Gensim library by Rehurek and Sojka (2010) for LDA topic extraction.

$$log(\frac{fraud_i}{1-fraud_i}) = \alpha + \sum_j \beta_j topic_{i,j} \tag{1}$$

In contrast to the BERT model, where we use the first and last 512 tokens of the MD&A text, for the LDA model we use the entire MD&A text. Before implementing the LDA model, we pre-processed the texts by performing lemmatization and removing stopwords. We optimize the *number of topics* parameter for the LDA model using the validation set. Hoberg and Lewis (2017) found 71 as the optimum number and Brown et al. (2020) found 31 as the optimum number of topics in their accounting fraud detection study. To accommodate these numbers, we searched within an interval of 10 and 150 topics, and found that the optimum number of topics in our setting that maximizes the validation AUC is 78. We implement the LDA model with 78 topics and find that for the interval between the years 2000 and 2013, the average AUC is 0.720 and the average NDCG@k is 0.127. LDA model captures altogether 20 fraud samples in its top 1% predictions in 14 test years.

Next, we discuss the performance of the BERT model based on text data. We use NVIDIA TESLA P100 GPU[6] for fine-tuning the BERT models. As discussed in Section 4.1, we optimize batch size and number of epochs using the validation set. Due to the computational constraints, we limit our search space for both the number of epochs and batch size within the set {1,2,..,8}. We find that a batch size of 8 and 3 epochs maximizes the validation AUC for $BERT_{final}$. In our validation set, we find that while $pred_{first}$ and $pred_{last}$ produces AUC values of 0.85 and 0.831 respectively, their rank average $pred_{final}$ produces an AUC value of 0.875. Moreover, we find that the rank correlation between $pred_{first}$ and $pred_{last}$ in the validation set is as low as 0.191 which indicates the degree of information diversity obtained by learning from these two models. This also supports our claim from the Section 4.1, where we conclude that the first and last parts of the MD&A section contain different information, and should therefore both be included.

We fine-tune both $BERT_{first}$ and $BERT_{last}$ with 3 epochs and a batch size of 8 to obtain predictions for each of our test years. We present the figures of ROC curves for BERT and LDA models in the Online supplement, as well as figures of NDCG as a function of deciles, which shows that LDA model does not catch up with BERT models. Generally, we find that NDCG scores for LDA model at different deciles are significantly smaller than those of BERT models. For example, from Table 2 for year 2002, we notice that all three BERT models capture 11 fraudulent observations (1% of total firms in the year 2002, see Table 1 Panel A) within their first 11 highest predicted scores for the test year 2002, thus achieving an NDCG@k score of 1. On the other hand, the LDA model cannot capture any fraudulent observations in the same top 11 predictions, thereby obtaining an NDCG@k score of 0. Table 2 contains the results on the yearly performance of both models on the textual data. We find that the average AUC is the highest for $BERT_{final}$, and also 15% higher than the average AUC of the LDA model. $BERT_{final}$ also achieves an average NDCG@k score of 0.675 which is 5.3 times more than that of LDA. While $BERT_{first}$ and $BERT_{last}$ captured altogether 100 and 97 fraudulent observations, respectively, within their top 1% predictions in the 14 test years, $BERT_{final}$ captured 107 fraudulent observations, which is 5.3 times more than that of LDA model, thus providing evidence of the economic significance of our model over textual benchmark model from the literature.

Table 2 also shows the yearly performance of LDA, $BERT_{first}$, $BERT_{last}$ and $BERT_{final}$ model. We observe that the LDA model is significantly under-performing across all test years with respect to other models. Interestingly, we find that in some years $BERT_{first}$ or $BERT_{last}$ have higher AUC scores than $BERT_{final}$, such as in the year 2002, despite the improvement in the validation year 1999. However, we notice that for 11 out of 14 test years, $BERT_{final}$ produces higher NDCG@k scores and for 10 out of 14 test years it captures a higher number of fraudulent observations in the top 1% prediction. Hence, we proceed with $BERT_{final}$ as it generalizes more across the years.

To statistically test whether the differences in average performance of other models comparing to $BERT_{final}$, we conduct an analysis of variance. The results (untabulated) yield p-values of less than 0.001, thereby also confirming the superiority of $BERT_{final}$ model in relation to other models.

5.2. Addressing RQ2

To examine how contextual information supplements the information obtained from existing quantitative methods, we first compare predictions of $BERT_{final}$ and RUSBoost model using the ensemble data. Then, we construct the Ensemble model of their predictions using rank average to understand the degree of complementarity.

It is important to note that in the ensemble data, the set of firm-year observations is a subset of the text data, as presented in Table 1.

[6] We are thankful to the Kaggle community for providing free access to GPU. Details can be found under the following link: https://www.kaggle.com/docs/efficient-gpu-usage.

To ensure that the training data size for each model is maximized and the algorithm's performance is improved, we extracted the predictions obtained from $BERT_{final}$ for the corresponding firm-year observations that are present in the ensemble data. This approach allows us to retain the maximum available training data for each model and avoid compromising the model's performance. Additionally, to further verify the robustness of our approach, we conducted another test as detailed in Section 7.5. In this test, we retrained the BERT models exclusively on the texts of the MD&As of firms that are present in the ensemble data. We found that this approach did not significantly alter the results, indicating that our approach is reliable.

Following Bao et al. (2020), we implement RUSBoost model from imbalance-learn library Lemaître et al. (2017) using 28 raw features as our independent variables. For the RUSBoost model, we optimize the number of trees using the validation set and find the following set of parameters that maximizes the RUSBoost AUC in the validation set. The final number of trees is set to be 2,500, the learning rate is 0.1, and we sample the same number of fraudulent and non-fraudulent observations during each iteration of the model following the approach of Bao et al. (2020). We employ the balancing strategy solely during the model training phase and not when generating predictions from the model. Specifically, we balance the 1994–1998 data for parameter optimization on the validation set, and we do not balance the 1999 data. Similarly, we balance only the training sets and not the test sets. For instance, to obtain predictions for the year 2000, we balance only the training set from 1995–1999 and not the test set for the year 2000, and so on.

With this optimized set of parameters, we train the models using the same procedure as previously described in $BERT_{final}$ model using ensemble data to obtain predictions for 14 test years. Finally, to produce the Ensemble model, we investigate the degree of complementarity that $BERT_{final}$ and the RUSBoost model share. We combine the predictions of those two models and check for the overall improvement. We take the weighted rank average of these two models' prediction values for each year to obtain the Ensemble prediction. $Ensemble_{pred} = w * \ rank(pred_{final}) + (1 - w) * \ rank(pred_{RUSBoost})$. Searching from the set {0.1,0.2,…,0.9}, we find 0.5 to be the optimum value of w using the validation set that produces maximum AUC on 1999 based on $Ensemble_{pred}$. Implementing such rank average of these two models results in the Ensemble model.

We present the results on the yearly performance of RUSBoost, $BERT_{final}$ and Ensemble model on the ensemble data in Table 3. RUSBoost model obtained an average AUC of 0.760 and an average NDCG@k score of 0.279 in the 14 test years. It captured in total 32 fraud observations in the top 1% predictions. The average AUC and NDCG@k of $BERT_{final}$ model in the ensemble data is 0.852 and 0.684 respectively, and it captures 96 fraud observations in the top 1% predictions altogether in 14 test years. This shows that the $BERT_{final}$ model outperforms the RUSBoost model by 12% when observing the AUC and captures three times more fraud observations in the top 1% predictions. $BERT_{final}$ also obtains 2.4 times more NDCG@k score than the RUSBoost model. The Ensemble model obtains an average AUC of 0.847 and an average NDCG@k of 0.550 and it captured 75 fraud observations in the top 1% predictions. We present the figures of ROC curves for $BERT_{final}$, RUSBoost, and Ensemble models in the Online supplement, as well as figures of NDCG as a function of deciles, which shows that RUSBoost model does not catch up with $BERT_{final}$ model. Generally, we find that NDCG scores for RUSBoost model at different deciles are significantly smaller than those of $BERT_{final}$ model.

It is visible from the results that the Ensemble model obtains competitive AUC scores with respect to $BERT_{final}$, however, the economic significance of $BERT_{final}$ is higher than the significance of the Ensemble model. This shows that combining $BERT_{final}$ and RUSBoost improves the AUC score of the standalone RUSBoost model by 11%, NDCG@k score by 2 times, and captures 2.34 times more fraudulent observations in the top 1% prediction. Interestingly, the Ensemble model could not outperform the $BERT_{final}$'s performance, likely because the lower performance of RUSBoost reduces the performance of the final Ensemble model in the rank average. Although the Ensemble model achieves competitive AUC to that of $BERT_{final}$, it captures 22% less fraudulent observations and achieves a 24% lower NDCG@k score. Even though the combination of the contextual and quantitative learning results in incremental improvements over the quantitative model alone, we finally conclude that the performance of the standalone contextual learning model is nevertheless higher.

We follow a similar approach as in RQ1 to assess the relative efficacy of our models. We conduct an analysis of variance to determine the differences in their average performance. The results (untabulated) indicate that the test obtains a p-value of less than 0.001, thereby confirming the superiority of Ensemble model over RUSBoost. However, when we compare the average AUC between the $BERT_{final}$ and Ensemble models, the result yields a p-value of more than 0.01, suggesting that the difference in their performance is not statistically significant.

6. Practical insights for financial investigators

In this section, we aim to provide further insights into how financial investigators can support their decisions based on the predictions obtained from $BERT_{final}$ model.

6.1. Insights using text data

We find that the MD&A writing style and the choice of words changes dynamically over time. Furthermore, we find evidence that firms with fraudulent filings tend to use more positive words and refrain from using negative words, possibly to disguise fraud. We further analyse the relative frequency of selection of positive and negative words from Loughran and McDonald's dictionary (Loughran and McDonald, 2011) in the MD&A reports separately for fraudulent and non-fraudulent filings across our training periods, and show that the use of positive words such as *gain*, *advances*, and *improvement* significantly increased over the years among the fraudulent filings, whereas the use of negative words such as *delays*, *excluding*, and *adverse* decreased over the years. We present example figures in the Online supplement.

Additionally, we investigate how the $BERT_{final}$ model, developed on validation data, performs also on later test years in order to show whether the writing style evolved.[7] We find that performance decreased significantly over years which indicates that the fraudulent firms adapt their writing style with time to produce misleading reports. This highlights the need for financial investigators to thoroughly scrutinize the signals that the model is picking to support their final decisions in prioritizing investigations.

6.2. Insights from financial data

In order to provide insights from financial data that can be used by financial investigators, we conduct two analyses. First, we analyse which firms are easy and difficult to identify for our $BERT_{final}$ model by uncovering the factors that are driving the probability of (mis)classifications. We consider the fraudulent observations which are correctly identified as within the top 1% predictions of the model and the non-fraudulent observations which are not in the top 1% predictions to be easy to identify. Reversely, the fraudulent observations that are not in the top 1% prediction and the non-fraudulent observations which are in the top 1% predictions, we consider to be difficult to identify. In our sample, we found altogether 21,691 firms that are easy to identify and 271 firms that are difficult to identify. Next, we use a decision tree classifier, along with the financial features from the ensemble data, to extract the rules that help to improve the understanding when a firm is difficult for $BERT_{final}$ to identify. We find that firms with higher inventories and higher annual sales are difficult for our model to identify, as well as the firms with higher inventories, lower annual sales and lower interest related expenses.

In the second analysis, we examine what factors drive a fraudulent 10-K filing to be erroneously identified by the $BERT_{final}$ model as a non-fraudulent filing and vice versa by performing a similar analysis on the set of firms that are difficult to identify. For that purpose, we consider firms which are erroneously identified by the model as fraudulent observation (firms identified by the $BERT_{final}$ model as belonging to top 1% predictions, but are in reality non-fraudulent), and the observations which are erroneously identified by the model as non-fraudulent (not in the top 1% prediction, but in reality fraudulent). We concatenate these two subsets for all 14 years of the test period and produce the data that have been difficult to identify by $BERT_{final}$ model. Set of all difficult firm-year observations produces altogether 143 fraud filings and 128 non-fraud filings. Next, we use a decision tree classifier and the financial features in order to understand what factors influence the erroneous classifications. We identify that fraudulent observations in firms with higher annual sales, lower net income, and higher annual close price have erroneously been identified by $BERT_{final}$ model as not fraudulent, whereas non-fraudulent observations in firms with low sales have been wrongly identified by the model as high-risk filings. These inferences can be further applied in making decisions by the financial investigators to prioritize their investigations. Investigators can use the first analysis to understand which firms could be difficult for the model to identify, and then on that set that is difficult to identify, the second analysis can be performed to better understand how to prioritize the investigations. Both analyses described above are presented in Fig. 4.

6.3. Enhancing interpretability with visual representations

Recent academic literature has shown a growing interest in enhancing the interpretability of advanced NLP models through visual representations. In this context, two instrumental tools have emerged: LIME (Local Interpretable Model-Agnostic Explanations) and BertViz. In Fig. 5(a) and 5(b), we provide illustrative examples of LIME and BertViz, respectively, to exemplify their application in this domain.

LIME (Ribeiro et al., 2016) is a model-agnostic tool that aids in the interpretation of complex NLP models, particularly in contexts like fraud detection. It functions by highlighting the significance of individual words within text data, allowing auditors to identify words contributing to higher or lower risk assessments in fraudulent activities. Fig. 5(a) illustrates an application of the LIME model to a fraudulent MD&A report, showcasing the words' importance in model predictions. In this visualization, blue words signify less risky terms, while red words highlight riskier elements. It is important to note that the company name is replaced with the term "company" for confidentiality.

On the other hand, BertViz (Vig, 2019) is designed specifically to facilitate the understanding of attention mechanisms within transformer-based models like BERT, which are commonly used in NLP tasks. These attention mechanisms enable models to allocate varying levels of importance to different words within a sentence when making predictions. BertViz provides a means to comprehend how much attention each word in a sentence pays to other words, shedding light on the model's decision-making process. For instance, Fig. 5(b) presents a practical example from our 10-K sample to illustrate how BERT operates in financial text and where it directs its attention. In the visualization, darker connections indicate a stronger focus on relevant words. In the first sentence of the sample, we observe that in the case of the word "we" the model focuses attention mostly on the word "have", the word "significant" attends to words "incurred" and "losses", and "since" attends to "inception". Besides providing insight into specific patterns of attention, we also observe that the model is able to capture linguistic notions, such as adjectives attending to corresponding nouns, and prepositions attending to their objects.

These visual representations, offered by LIME and BertViz, enable auditors and analysts to grasp the inner workings of complex NLP models, allowing them to make more informed decisions and gain deeper insights into textual data for tasks like fraud detection and

[7] $BERT_{final}$ model is developed on the period from 1994 to 1998, and the AUC for the test year 2000 is 0.738, for 2005 is 0.630, for 2010 is 0.609, and for 2013 is 0.522.

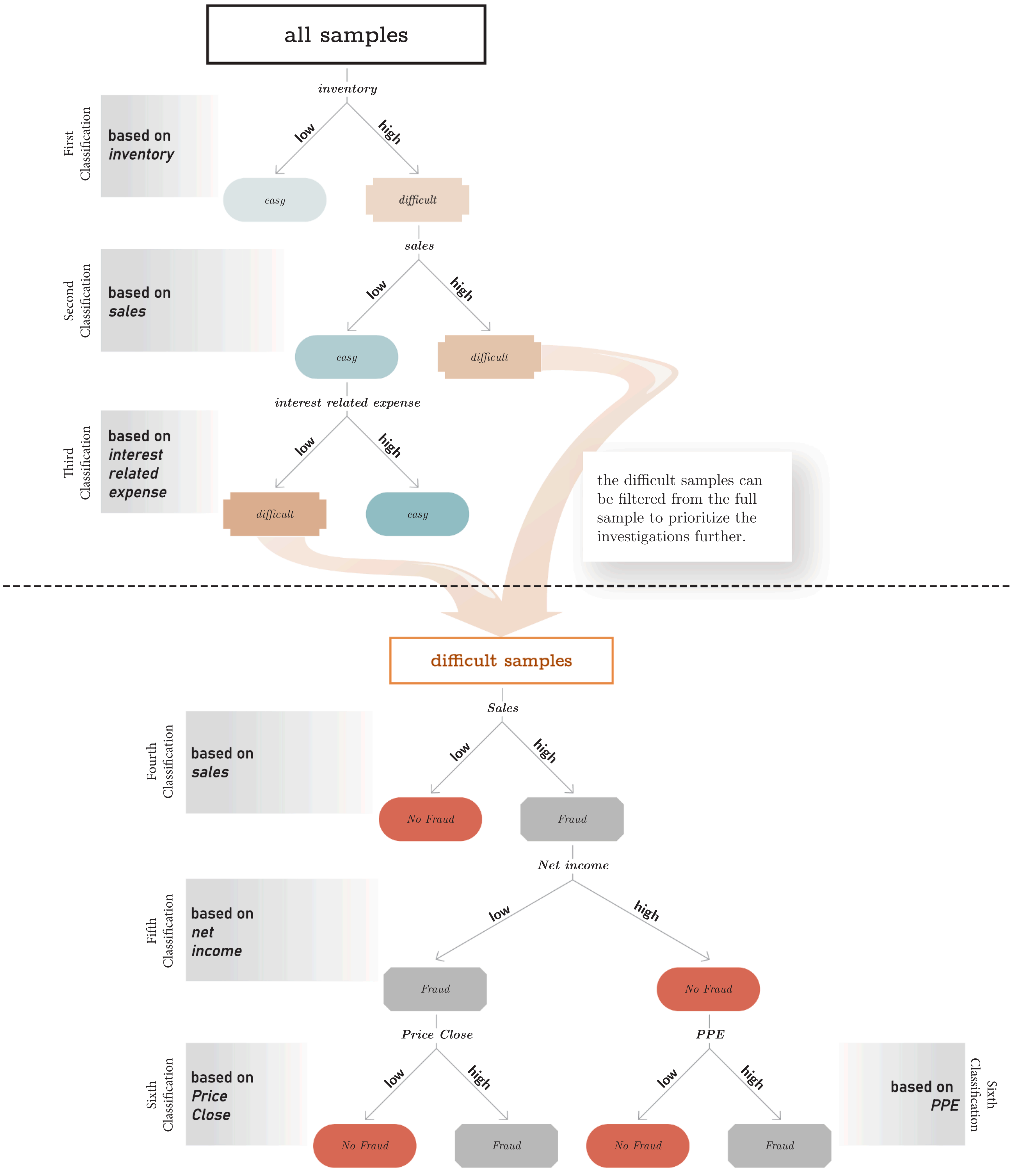


Fig. 4. Two analyses that are providing insights from financial data. First, we analyse which firms are difficult to identify for our BERT model. Next, we examine what drives fraudulent observations to be erroneously identified by the BERT model as non-fraudulent observations and vice versa by a similar analysis on the set of difficult filings.

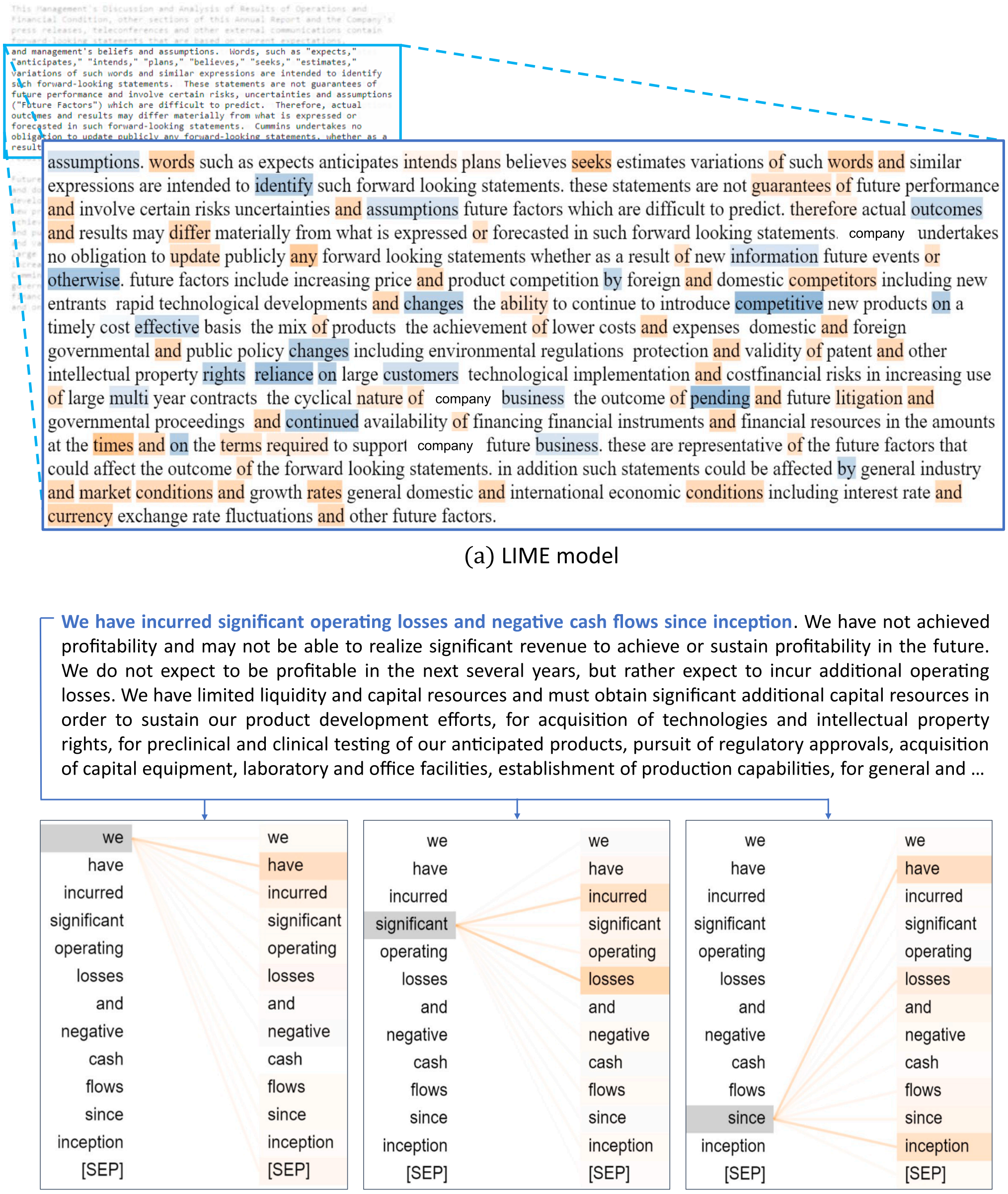
financial analysis.

7. Supplementary analysis

To further increase the confidence in our findings, we conduct the following supplementary analysis.

7.1. Tackling class imbalance

The accounting fraud detection problem suffers from class imbalance, since the average percentage of fraudulent 10-K filings in our text data is only 1.30%. Therefore, we consider whether accounting for class imbalance would help in improving the baseline $BERT_{final}$ model. We attempt to tackle class imbalance by specifying class weights in our fine-tuning procedure using validation set, which would cause the model to pay more attention to the fraud observations. We observe that assigning class weights by obtaining them from the training set does not improve predictions on the validation set. We use different combinations of batch size and number of epochs from the set {1,2,…,8} and specify class weights of fraud and no-fraud observations in the validation set. We find that the $BERT_{final}$ model



This Management's Discussion and Analysis of Results of Operations and Financial Condition, other sections of this Annual Report and the Company's press releases, teleconferences and other external communications contain forward-looking statements that are based on current expectations, and management's beliefs and assumptions. Words, such as "expects," "anticipates," "intends," "plans," "believes," "seeks," "estimates," variations of such words and similar expressions are intended to identify such forward-looking statements. These statements are not guarantees of future performance and involve certain risks, uncertainties and assumptions ("Future Factors") which are difficult to predict. Therefore, actual outcomes and results may differ materially from what is expressed or forecasted in such forward-looking statements. Cummins undertakes no obligation to update publicly any forward-looking statements, whether as a result

assumptions. words such as expects anticipates intends plans believes seeks estimates variations of such words and similar expressions are intended to identify such forward looking statements. these statements are not guarantees of future performance and involve certain risks uncertainties and assumptions future factors which are difficult to predict. therefore actual outcomes and results may differ materially from what is expressed or forecasted in such forward looking statements. company undertakes no obligation to update publicly any forward looking statements whether as a result of new information future events or otherwise. future factors include increasing price and product competition by foreign and domestic competitors including new entrants rapid technological developments and changes the ability to continue to introduce competitive new products on a timely cost effective basis the mix of products the achievement of lower costs and expenses domestic and foreign governmental and public policy changes including environmental regulations protection and validity of patent and other intellectual property rights reliance on large customers technological implementation and costfinancial risks in increasing use of large multi year contracts the cyclical nature of company business the outcome of pending and future litigation and governmental proceedings and continued availability of financing financial instruments and financial resources in the amounts at the times and on the terms required to support company future business. these are representative of the future factors that could affect the outcome of the forward looking statements. in addition such statements could be affected by general industry and market conditions and growth rates general domestic and international economic conditions including interest rate and currency exchange rate fluctuations and other future factors.

(a) LIME model

We have incurred significant operating losses and negative cash flows since inception. We have not achieved profitability and may not be able to realize significant revenue to achieve or sustain profitability in the future. We do not expect to be profitable in the next several years, but rather expect to incur additional operating losses. We have limited liquidity and capital resources and must obtain significant additional capital resources in order to sustain our product development efforts, for acquisition of technologies and intellectual property rights, for preclinical and clinical testing of our anticipated products, pursuit of regulatory approvals, acquisition of capital equipment, laboratory and office facilities, establishment of production capabilities, for general and …

(b) BertViz model

Fig. 5. Two examples of models that can assist the auditors in identifying and visually presenting important words in the prediction process. (a) shows the implementation of the LIME model on a fraudulent MD&A report. (b) demonstrates the application of BertViz to analyze a specific sentence from the example MD&A report.

obtains the highest validation AUC of 0.856 with 4 epochs and batch size 4 after specifying class weights. This score is higher without accounting for class imbalance (0.875), as described in Section 5.1.

7.2. Robustness check

We perform an additional robustness check following research conducted by Siano and Wysocki (2021). For every test year, we first identify the 30 most frequent words in the corresponding training set and replace these words from the test set with a random word *wxyz* in our text data. We find that replacing these words does not affect the model extensively as it produces an average AUC of 0.824, is able to produce an average NDCG@k score of 0.637, and captures 94 fraud observations altogether in the 14 test years. This is lower than the values obtained with $BERT_{final}$ model, where AUC is 0.826, NDCG@k score is 0.675, and it captures 107 fraud observations. It shows that our model is not extensively dependent on the most frequent words and even after replacing them with a random word, it can retain its contextual learning from the texts.

7.3. Excluding serial frauds

In some firms serial fraudulent behavior is detected, where fraud spans over multiple consecutive years. On the other hand, it is also likely that the same professional body is responsible for filling the 10-K reports. Therefore, it is important to investigate whether our model is gaining knowledge about fraudulent behavior, or merely learning the writing pattern from serial fraud filings. For each test year, we first identify fraudulent firm-year observations in the training set which are appearing multiple times. Next, we keep only one fraudulent observation of such a serial fraud firm and remove other observations. Specifically, we keep the first fraudulent year's observation and remove the consecutive fraudulent observations.[8]

This results in a drop of 51% of total fraud observations. However, even after such a significant drop of total fraudulent observations in the training set, retraining the $BERT_{final}$ upon excluding serial fraud did not reduce its performance significantly. The retrained model finally produces an average AUC of 0.763 (comparing to 0.826 for $BERT_{final}$), captures 72 fraud samples altogether in the 14 test years (comparing to 107 for $BERT_{final}$), and an average NDCG@k of 0.508 (comparing to 0.675 for $BERT_{final}$). The drop in performance can be accounted for by the considerable drop in fraudulent observations in the training data. This analysis shows that the model is in general learning about the inherent nature of fraud and not only picking up the style and language of serial fraud firms, which also demonstrates the possibility of generalizing the approach to a broader population.

7.4. Ensembling all models

We investigate whether ensembling $BERT_{final}$, LDA and the RUSBoost model would result in further improvement. For this experiment, we first compute weighted average of predictions from these 3 models as $Ensemble_{all} = w_1 * rank(pred_{final}) + w_2 * rank(pred_{LDA}) + (1 - w_1 - w_2) * rank(pred_{RUSBoost})$, where $w_1, w_2 \in (0, 1)$. We search the optimum values of w_1 and w_2 using validation set from the search space of {0.1,0.2,…,0.9}, and find that $(w_1, w_2) = (0.3, 0.4)$ maximizes the validation AUC on the year 1999. With these values of w_1 and w_2, we compute the $Ensemble_{all}$ for all the 14 years using ensemble data. We find that $Ensemble_{all}$ produces an average AUC of 0.836, an average NDCG@K score of 0.502, and it captures altogether 69 fraudulent observations in the 14 test years. Although we show that ensembling all three models with a simple weighted average does not seem to improve performance of the model, the future research could use more sophisticated methods such as bagging and boosting that could potentially further improve performance.

7.5. Retraining BERT on ensemble data subset

To ensure the robustness of our results, we conducted a separate analysis where we trained the BERT model solely on the subsample of text data that is present in the ensemble data. In other words, we trained the BERT models on the texts of 10-K reports present in the ensemble data and evaluated RQ2 on this subset of data. The detailed results of this analysis are included in the Online Appendix.

Our findings showed that the results of RQ2 were statistically comparable and not significantly different whether we trained BERT models on the entire text data or only on the subset of data present in the ensemble. The average AUC and NDCG@K scores of the $BERT_{final}$ model were 0.842 and 0.685, respectively, compared to 0.852 and 0.684, respectively, when BERT was trained on the entire text data. $BERT_{final}$ captured a total of 97 fraudulent observations in its top 1% predictions, compared to 96 when BERT was trained on the entire text data.

Similarly, the ensemble model produced average AUC and NDCG@k scores of 0.839 and 0.541, respectively, compared to 0.847 and 0.550, respectively, when BERT was trained on the entire text data. The ensemble model captured a total of 72 fraudulent observations in its top 1% predictions, compared to 75 when BERT was trained on the entire text data.

We conducted two-sided t-tests, which resulted in p-values of more than 0.1 when comparing $BERT_{final}$'s and the ensemble model's predictions between Tables 3 and 4. This indicates that the results are statistically similar whether we train the BERT model with the full text data or only with the text of firms present in the ensemble data.

8. Discussion and conclusion

Because of the progress in the natural language processing models in the past years, the simple financial information extracted from the balance sheets is no longer enough to supplement the textual models. This is also visible from our results, presented in Table 3, where the performance of the final BERT model is similar to the performance of the Ensemble model, and even outperforms it in terms of economic significance. Instead of using raw financial information, more complex financial measures could be used to improve the prediction.

In the overview table of our sample (Table 1), it is visible that the percentage of detected fraudulent 10-K filings is declining over

[8] For example, for a specific firm and for test year t, if fraud is detected in three years in the training period ($t-4, t-3$, and $t-2$). We then keep only the first fraudulent observation (in $t-4$), and exclude following fraudulent observations (in years $t-3$ and $t-2$).

time. This potentially indicates that an increasing number of fraudulent filings are being undetected. Incidentally, SEC indicated that their focus shifted during the financial crisis in 2008, focusing more on the collateralized debt obligation (CDOs), residential mortgage-backed securities (RMBS), and Ponzi schemes (Ceresney, 2013). The change of focus could also explain the declining number of detected fraudulent filings in our sample. Moreover, some studies suggest that the number of fraudulent filings is significantly higher than what is actually detected, some citing as much as 11% of the large U.S. public corporations allegedly committing fraud (Dyck et al., 2021). Working with regulators and incorporating state-of-the-art tools from natural language processing could help detect more fraudulent observations, beyond the currently detected ones.

As with any research endeavor, our study is not immune to certain limitations that are worth noting. Firstly, we relied on a dated AAER dataset, which restricted our analysis of the detection of fraudulent 10-K filings to the year 2013. Given the dynamic nature of financial fraud, it would be interesting to investigate how our model performs in more recent years. Secondly, we acknowledge that there are other ensembling methods, such as boosting or stacking, that could be explored to further enhance the model's performance. However, due to the associated computational challenges, we did not explore them in this study. It could potentially be interesting for the future research to explore these options in an efficient way. Thirdly, an additional limitation of this study pertains to the consideration of firm size. Specifically, we do not differentiate between firms based on their size. In cases where the firm with the highest probability of being fraudulent is relatively small in size, its detection may not hold significant economic or social relevance. This underscores the need for future research to create a dependent variable quantifying economic impacts and to address these nuances for a more comprehensive analysis of accounting fraud.

We believe that the potential of contextual language learning in detecting accounting frauds is vast and that a lot of facets are still left unexplored. In the following, we provide some ideas for future research. For example, future research can be carried out to explore if extracting a deeper sense of the business text (from financial reports, conference calls, or corporate social responsibility reports) can help in forecasting companies' earnings or in predicting audit quality. Another interesting direction for future research could be contextual learning based on academic and professional publications, such as earnings announcements, earnings call transcripts, analysts' reports, and journals. Researchers can develop an accounting BERT model by pre-training on publications from accounting literature and further direct it in order to tackle domain-specific tasks.

The problem of accounting fraud detection sparks interest among auditors, investors, and researchers. However, solving this problem is not easy, and detecting fraud is neither easy nor free. Previous literature mostly explored the potential of using different quantitative features (such as information from financial statements or stock market) to detect the likelihood of fraud, and recent literature started investigating the use of textual analysis to detect fraud. We build on this research and show how including context from financial reports helps in detecting accounting fraud. We apply the BERT model to the accounting field to learn the contexts of the MD&A section of annual 10-K reports and further direct that contextual knowledge to detect accounting fraud. We find that the BERT model significantly outperforms previously used textual and quantitative models. Moreover, we find that our final model identifies five times more fraudulent observations than the textual benchmark by investigating the same number of observations, and three times more than the quantitative benchmark.

Dataset

Dechow, Ge, Larson and Sloan (2011) Predicting Material Accounting Misstatements. Contemporary Accounting Research, 28: 17–82. https://doi.org/10.1111/j.1911–3846.2010.01041.x, dataset can be obtained from: https://sites.google.com/usc.edu/aaerdataset/home.

U.S. Securities and Exchange Commission (SEC) EDGAR database, https://www.sec.gov/edgar/searchedgar/companysearch.html.

Compustat, https://wrds-www.wharton.upenn.edu/.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

The information on MD&A section of annual 10-K reports is publicly available from the SEC EDGAR database. The accounting and financial data is downloaded from Compustat. The information on misstatements comes from the SECs Accounting and Auditing Enforcement Releases (AAERs), as compiled by Dechow, Ge, Larson and Sloan (2011) dataset. The code will be made available on request.

Declaration of competing interest

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

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.accinf.2024.100682.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., 2016. TensorFlow: A system for Large-Scale machine learning, in: 12th USENIX symposium on operating systems design and implementation (OSDI 16), pp. 265–283.

Alberti, C., Lee, K., Collins, M., 2019. A bert baseline for the natural questions. arXiv:1901.08634.

Bao, Y., Ke, B., Li, B., Yu, Y.J., Zhang, J., 2020. Detecting accounting fraud in publicly traded us firms using a machine learning approach. J. Acc. Res. 58, 199–235.

Berkin, A., Aerts, W., Van Caneghem, T., 2023. Feasibility analysis of machine learning for performance-related attributional statements. Int. J. Acc. Inf. Syst. 48, 100597.

Berns, J., Bick, P., Flugum, R., Houston, R., 2022. Do changes in md&a section tone predict investment behavior? Financial Rev.

Brown, N.C., Crowley, R.M., Elliott, W.B., 2020. What are you saying? Using topic to detect financial misreporting. J. Account. Res. 58, 237–291.

Bushee, B.J., Gow, I.D., Taylor, D.J., 2018. Linguistic complexity in firm disclosures: obfuscation or information? J. Acc. Res. 56, 85–121.

Cecchini, M., Aytug, H., Koehler, G.J., Pathak, P., 2010. Detecting management fraud in public companies. Manage. Sci. 56, 1146–1160.

Cecchini, M., Aytug, H., Koehler, G.J., Pathak, P., 2010. Making words work: using financial text as a predictor of financial events. Decis. Support Syst. 50, 164–175.

Ceresney, A., 2013. Sec.gov — financial reporting and accounting fraud URL: https://www.sec.gov/news/speech/spch091913ac. (Accessed on 05/11/2021).

Craja, P., Kim, A., Lessmann, S., 2020. Deep learning for detecting financial statement fraud. Decis. Support Syst. 139, 113421.

Dechow, P.M., Ge, W., Larson, C.R., Sloan, R.G., 2011. Predicting material accounting misstatements. Contemp. Acc. Res. 28, 17–82.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. Proceedings of the 2019 Conference of the North.1 doi:10.18653/v1/n19-1423.

Durnev, A., Mangen, C., 2020. The spillover effects of md&a disclosures for real investment: the role of industry competition. J. Acc. Econ. 70, 101299.

Dyck, A., Morse, A., Zingales, L., 2010. Who blows the whistle on corporate fraud? J Finance 65, 2213–2253.

Dyck, I., Morse, A., Zingales, L., 2021. How pervasive is corporate fraud? Rotman School of Management Working Paper.

Feldman, R., Govindaraj, S., Livnat, J., Segal, B., 2010. Management's tone change, post earnings announcement drift and accruals. Rev. Acc. Stud. 15, 915–953.

Glancy, F.H., Yadav, S.B., 2011. A computational model for financial reporting fraud detection. Decis. Support Syst. 50, 595–601.

Goel, S., Gangolly, J., 2012. Beyond the numbers: mining the annual reports for hidden cues indicative of financial statement fraud. Intell. Syst. Acc. Finance Manage. 19, 75–89.

Goel, S., Gangolly, J., Faerman, S.R., Uzuner, O., 2010. Can linguistic predictors detect fraudulent financial filings? J. Emerg. Technol. Acc. 7, 25–46.

Hoberg, G., Lewis, C., 2017. Do fraudulent firms produce abnormal disclosure? J. Corporate Finance 43, 58–85.

Holder-Webb, L., Cohen, J.R., 2007. The association between disclosure, distress, and failure. J. Bus. Ethics 75, 301–314.

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. Decis. Support Syst. 50, 585–594.

Karpoff, J.M., Koester, A., Lee, D.S., Martin, G.S., 2017. Proxies and databases in financial misconduct research. Acc. Rev. 92, 129–163.

Larcker, D.F., Zakolyukina, A.A., 2012. Detecting deceptive discussions in conference calls. J. Acc. Res. 50, 495–540.

Lemaı̂tre, G., Nogueira, F., Aridas, C.K., 2017. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. 18, 1–5. URL: http://jmlr.org/papers/v18/16-365.

Li, F., 2008. Annual report readability, current earnings, and earnings persistence. J. Acc. Econ. 45, 221–247.

Liu, W.P., Yen, M.F., Wu, T.Y., 2022. Report users' perceived sentiments of key audit matters and firm performance: evidence from a deep learning-based natural language processing approach. J. Inf. Syst. 36, 191–209.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. J. Finance 66, 35–65.

Loughran, T., McDonald, B., 2016. Textual analysis in accounting and finance: a survey. J. Acc. Res. 54, 1187–1230.

Minhas, S., Hussain, A., 2016. From spin to swindle: identifying falsification in financial text. Cogn. Comput. 8, 729–745.

Moffit, K., Burns, M., Felix, W., Burgoon, J., 2010. Using lexical bundles to discriminate between fraudulent and non-fraudulent financial reports on. SIG-ASYS Pre-ICIS 2010 workshop.

Muslu, V., Radhakrishnan, S., Subramanyam, K., Lim, D., 2015. Forward-looking md&a disclosures and the information environment. Manage. Sci. 61, 931–948.

Pratt, G.A., 2015. Is a cambrian explosion coming for robotics? J. Econ. Perspect. 29, 51–60.

Purda, L., Skillicorn, D., 2015. Accounting variables, deception, and a bag of words: assessing the tools of fraud detection. Contemp. Acc. Res. 32, 1193–1223.

Rehurek, R., Sojka, P., 2010. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks, Citeseer.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. why should i trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144.

Rogers, J.L., Van Buskirk, A., Zechman, S.L., 2011. Disclosure tone and shareholder litigation. Acc. Rev. 86, 2155–2183.

Securities Exchange Commission, 2003. Interpretation: Commission guidance regarding management's discussion and analysis of financial condition and results of operations. Securities Act Release 34–48960.

Siano, F., Wysocki, P., 2021. Transfer learning and textual analysis of accounting disclosures: applying big data methods to small (er) datasets. Acc. Horizons 35, 217–244.

Sun, C., Qiu, X., Xu, Y., Huang, X., 2019. How to fine-tune bert for text classification? China National Conference on Chinese Computational Linguistics. Springer, pp. 194–206.

Vig, J., 2019. A multiscale visualization of attention in the transformer model. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. https://doi.org/10.18653/v1/p19-3007.

Yang, J., Wang, M., Zhou, H., Zhao, C., Zhang, W., Yu, Y., Li, L., 2020. Towards making the most of bert in neural machine translation. In: Proceedings of the AAAI conference on artificial intelligence, pp. 9378–9385.

Yang, W., Xie, Y., Lin, A., Li, X., Tan, L., Xiong, K., Li, M., Lin, J., 2019. End-to-end open-domain question answering with. In: Proceedings of the 2019 Conference of the North. https://doi.org/10.18653/v1/n19-4013.

Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li, H., Liu, T., 2020. Incorporating bert into neural machine translation. In: International Conference on Learning Representations.