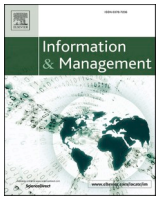
Information & Management 62 (2025) 104055

Contents lists available at ScienceDirect

Information & Management

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



Sentiment-devoid lexicons: A novel method for domain-specific textual analysis in business and governance documents

Wentao Ma [a], Shuk Ying Ho [b,*]

[a] International Business School, Xi'an Jiaotong-Liverpool University, China
[b] Research School of Accounting, The Australian National University, Australia

ARTICLE INFO

Keywords:
Textual analysis
Dictionary
Sentiment-devoid
SEC reviews
Comment letters
Information technology control weaknesses

ABSTRACT

Our study proposes and tests a method for developing domain-specific dictionaries tailored for textual analysis in information systems research. Traditionally, dictionaries have been widely used for content classification according to sentiment; however, we introduce an alternative approach focused on creating dictionaries from sentiment-devoid documents. We demonstrate this method by developing a dictionary specific to Securities and Exchange Commission (SEC) investigations. Analyzing 150,432 publicly available SEC documents, we gained insights into the semantics of communications between the SEC and firms. To evaluate the dictionary, we analyzed SEC comment letters to predict the likelihood of firms reporting information technology control weaknesses (ITCWs), information technology audit fees, and cyber risks. Our dictionary outperformed five benchmarking dictionaries, explaining a higher proportion of variance in ITCW likelihood, information technology audit fees, and cyber risks. This study enhances the effectiveness of dictionaries in analyzing sentiment-devoid business and governance documents and results in a specialized dictionary for SEC communications.

1. Introduction

Given the increasing importance of information systems (IS) policies in corporate governance, researchers are analyzing disclosures published by governments and firms as data sources [1,2]. These sources provide rich information; for example, the Sarbanes–Oxley Act of 2002 requires management and auditors to regularly publish their assessment of firm information technology (IT) controls over enterprise systems (C. [3]). In 2022, the Securities and Exchange Commission (SEC) proposed new regulations requiring management to report material cybersecurity incidents, as well as technology risk management and governance, leading to a significant increase in the volume and variation of firm cybersecurity disclosures. These disclosures lack sentiment clues, such as those found in informal content (e.g., "like" and "good" in consumer reviews). In addition, the specificity of technology domains significantly influences the style and use of business language and keywords in these disclosures [4]. The language used in cybersecurity disclosures, for instance, differs from that used in IT control assessments because each domain is governed by distinct regulations and industry practices. This variability in language and context makes it challenging for one-size-fits-all, sentiment-based dictionaries to capture the specific meaning of such disclosures accurately. Instead, domain-specific dictionaries are needed to reflect the regulatory and business context, allowing researchers to more effectively analyze these disclosures and derive relevant insights for IS research.

Our study proposes a method for constructing dictionaries tailored to official disclosures that lack typical sentiment-based language. Unlike content influenced by stock market performance, such as earnings announcements, official disclosures—especially those from regulatory bodies and firms—rely on formal, sentiment-devoid language. This makes traditional sentiment-based analysis less effective in these contexts [5]. By focusing on the unique characteristics of such disclosures, our approach aims to enhance the accuracy and relevance of textual analysis in IS research, particularly in domains in which formal, sentiment-free documents serve as crucial data sources.

A dictionary consists of clusters of words, word stems, or phrases that

* Corresponding author.
E-mail addresses: wentao.ma@xjtlu.edu.cn (W. Ma), susanna.ho@anu.edu.au (S.Y. Ho).

https://doi.org/10.1016/j.im.2024.104055
Received 29 June 2023; Received in revised form 20 September 2024; Accepted 11 November 2024
Available online 16 November 2024

express similar emotions or attitudes, or discuss related topics within a given knowledge domain [6]. Such dictionaries support textual analysis in which computer programs systematically process words and phrases matched to the dictionary extracted from a document, determine the semantic orientation, and assign values to variables and measures of researchers' interests. These dictionaries enable researchers to scrutinize meanings embedded in complex and abundant sentences and paragraphs in documents. Sentences and paragraphs are unstructured data with no predefined data model and are not organized in a predefined manner [7]. Most dictionaries contain wordlists that enable researchers to classify sentimental words according to valence (positive, neutral, and negative). However, these dictionaries may not perform effectively when applied to specific problems in domains different from those in which they were developed, particularly in domains where professional bodies communicate using formal language and few sentimental words [8]. As a result, sentiment-based dictionaries fail to effectively model how information providers communicate their ideologies, thoughts, and experiences through text to information recipients [9]. Further, each type of business document has a unique focus. In this study, we concentrate on SEC disclosures as the primary source for developing our dictionary. Consequently, our constructed dictionary can be applied to examining issues that align with the SEC's interests. When the focus of analysis shifts, researchers may seek different source documents for dictionary construction. For example, researchers may consider systems compliance and integrity reports for technology systems security and disruptions, Form 8-Ks for announcing major events of which shareholders should be aware (e.g., cybersecurity incidents), and critical audit matter reports for revealing issues related to IT controls.

We suggest a dictionary-development approach and use this approach to build a dictionary tailored for the analysis of SEC evaluations concerning corporate disclosures. In the United States, the SEC promotes firms' compliance with reporting regulations and conducts routine investigations in which officials gather facts and evidence through informal inquiries and by reviewing firms' compliance with reporting regulations in financial statements, disclosures, brokerage records, and trading data.[1] We are particularly interested in the SEC review documents that outline firms' financial compliance problems and request firms' clarifications and supporting documents because these documents cover important business topics such as information technology control weaknesses (ITCWs; C. [3]; W. Y. [10]), data protection [11], and cybersecurity breaches [12] and are therefore interesting to IS researchers. Before 2004, the SEC review documents were accessible only to individuals or entities that filed a Freedom of Information Act request. At that time, researchers read accounting and auditing enforcement releases to discover the SEC's review foci. Since August 1, 2004, the SEC has filed corporate reports, including the SEC's comment letters (CLs) and corporate filers' responses, on the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR).[2] In a publication review by Cunningham and Leidner [13], 80 published (or working) papers considered the SEC review variables, and 90 % of these were published (or circulated) after 2016. Authors of these papers used counts of CLs and review durations to predict factors such as the quality of firms' enterprise and security systems [14], cybersecurity incidents [15], and auditors' cybersecurity risk assessments [16]. However, CL counts and review durations are correlated with problem scale, information structuredness, and complexity, as well as with the nature of the review. For example, a review of firms' revenue recognition takes an average of twice as many days as other reviews [14]. Therefore, abstracting analyses to the SEC review level may produce confounding results [4,13]. One seminal work that considers SEC review content is Ryans [4], who manually classified words and constructed variables for the naive Bayesian classification models.

In this paper, we scrutinized 150,432 CLs and learned the semantics of dialogues between the SEC and firms, and between the SEC and the public [17]. The resultant dictionary contained 934 words. To test the practicality of our dictionary, we used it to analyze the SEC's master letters, that is, the initiating letters of an SEC review, and to predict one important firm subsequent action—firm disclosure of ITCW—in the following year, possibly triggered by the SEC investigation. We also examined our dictionary's practicality in a risk-assessment scenario by using the dictionary to predict firms' IT audit fees and cyber risks in the following year. We then compared the performance of our dictionary with that of five other dictionaries that have been widely used in business research: the Loughran and McDonald [17] dictionary, the Harvard General Inquirer dictionary [18], the Corporate Finance Institute (CFI) financial dictionary [19], the Investopedia financial dictionary [20], and the Henry [21] dictionary.

Our study contributes to both theory and practice in several ways. First, we proposed a dictionary-construction method tailored to sentiment-devoid documents, incorporating neural network modeling to rank word importance, unlike traditional methods that rely on manual coding (e.g., [22]). This approach enables researchers to develop dictionaries specific to business contexts in which sentiment analysis is less effective in capturing writers' intentions [5]. Although limited dictionaries exist for sentiment-devoid texts, our method addresses the challenge of selecting insightful words in these contexts. By minimizing human judgment and facilitating replication, our approach supports the development of high-quality, domain-specific dictionaries, helping IS researchers analyze textual data quantitatively and qualitatively. This may lead to novel constructs that enhance the empirical validity of textual analysis in IS research. Second, we developed a comprehensive dictionary based on neural network modeling of a large sample of SEC CLs. This responds to Cunningham and Leidner's [13] call for systematic, human-minimized textual analysis of SEC CLs. The resulting dictionary will be available to researchers and practitioners, enabling more precise content analyses of SEC CLs. Third, we tested our dictionary in several IT-related contexts, including detecting ITCWs, predicting IT audit fees, and identifying cyber risks. Our results showed that documents that contain more words from our dictionary correspond to a higher likelihood of ITCWs, higher IT audit fees, and increased cyber risks. Our dictionary outperformed the widely used Loughran and McDonald dictionary, suggesting that IS and IT documents may be largely sentiment-devoid and require an approach not based on sentiment. This opens new research opportunities in IS and offers valuable insights for practitioners, such as chief information officers, in understanding and complying with IT regulations, particularly those enforced by the SEC.

2. Role of textual analysis in IS research

Textual analysis is gaining importance in IS research. In the era of big data, researchers have a strong interest in inferring individuals' preferences, attitudes, and opinions from their digital footprints (e.g., web logs and social media posts), which has led to the widespread use of sentiment analysis to examine these factors. The primary form of textual analysis is sentiment analysis [23]. One research area that extensively uses sentiment analysis is social media analytics, in which researchers process social media posts to predict stock market movements and firm performance. For example, Ho et al. [24] applied sentiment analysis to Twitter data to predict firms' sales growth, and Deng et al. [25] used StockTwits data to demonstrate a bidirectional interaction between tweet sentiment and stock returns. In healthcare IS, Dissanayake et al.

[1] Refer to https://www.sec.gov/divisions/corpfin/cffilingreview

[2] Before 2012, the SEC policy was to upload CLs to EDGAR no later than 45 days after the end of the review. Since January 1, 2012, the delay has been decreased to 20 business days.

[26] analyzed patients' self-descriptions from CrowdMed and found that strongly negative self-descriptions bias medical practitioners' judgments. Another area that applies sentiment analysis to big data is consumer reviews. For instance, Luo et al. [27] mined hundreds of thousands of expert blogs and found that expert blog sentiment on a focal brand negatively affects how consumers perceive competitors. Chang et al. [28] used advanced sentiment analysis on consumer reviews to infer customer dissatisfaction and identify areas for product and service improvement.

The quality of sentiment analysis depends on the quality of the dictionary used [6]. In sentiment analysis, computer programs count words in a text, match them with words in a sentiment dictionary, and return measures such as sentiment scores. Different forms of a word (e.g., "punish" and "punished," "lawsuit" and "lawsuits," "bad" and "worse") provide additional sentiment clues [29]. A verb in the past tense carries less intense sentiment than a verb in the present tense, and "worst" indicates a more negative outcome than "bad." High-quality dictionaries should be comprehensive and developed using minimal subjective interpretation but sufficient domain- or context-specific consideration [30,31]. These features minimize misclassification [17] and increase accuracy in sentiment measurement, enhancing explanatory power in estimations [6]. Well-known dictionaries for sentiment analysis include Bing, the Multi-Perspective Question Answering Opinion Corpus, the National Research Council Canada Emotion Lexicon, and the Quantitative Discourse Analysis Package.[3]

Researchers frequently use dictionaries such as the Harvard General Inquirer and Diction for textual analysis of business documents. The Harvard General Inquirer, one of the first dictionaries applied in business research, was originally developed for sociology and psychology [18]. It includes a collection of wordlists designed to measure over 100 document attributes.[4] This dictionary has been widely used to analyze news articles and firm annual reports. For example, X. Li et al. [32] demonstrated that using the Harvard General Inquirer to analyze news sentiment can predict stock price movements effectively.

Diction has 31 wordlists [33]. Each wordlist contains a small number of precisely classified words.[5] This enables researchers to construct dictionaries that meet their needs. For example, Davis and Tama-Sweet [34] used lists of words related to blame, hardship, and denial to construct a dictionary for analyzing pessimism in the "Management Discussion and Analysis" section of firms' annual reports and documented the strategic reporting behaviors of managers. Specifically, they found that managers avoid using pessimistic words related to low profitability in the future. However, given that the Harvard General Inquirer and Diction are not context-specific linguistic dictionaries, their classification accuracy decreases when applied to business documents [35].

Henry [21] and Loughran and McDonald [17] are sentiment-based dictionaries designed for analyzing business documents. The dictionary by Henry [21] was developed using earnings press releases for the telecommunications and computer services industries only. It contains two wordlists (positivity and negativity) and 290 words. Henry used this dictionary to process 1,366 earnings press releases and found a positive relationship between the sentiment of press releases and abnormal stock returns. However, one major limitation of Henry's dictionary is its short wordlist, which contains only 85 negative words and does not include some negative words frequently used in business documents, such as "loss," "losses," "adverse," and "impairment."

Loughran and McDonald's [17] dictionary is specific to formal and financial documents, specifically, firm annual reports. Loughran and McDonald aimed to identify words that have negative implications, which can effectively measure the tone of firm annual reports.[6] To do this, they considered widely used business words while excluding sentiment words that are not used or have different meanings in the context of annual reports. For example, their dictionary does not include "outstanding" (which is generally a positive word) because it has a different meaning in such a specific context.[7] Loughran and McDonald confirmed that their dictionary is highly effective when used to measure the sentiment of documents written by business professionals (e.g., managers and business journalists), such as business journals [36], to predict firm financial performance.

Notably, when applied to documents other than annual reports and articles in business journals, the quality of sentiment analyses using Loughran and McDonald's [17] dictionary was less than satisfactory. For example, Chen et al. [37] found limited predictability of stock returns for the proportion of words from the Loughran and McDonald [17] dictionary in articles and comments on Seeking Alpha.

In our dictionary development, we focused on SEC investigations. SEC investigations (also referred to as "SEC reviews") refer to the process in which the SEC examines various documents, filings, and activities related to securities offerings, public companies, and market participants. The SEC conducts its investigations to ensure firm compliance with securities laws and regulations, promote transparency, and protect investors. The review documents convey high-value information, and each review focuses on different problems. This means that any textual analyses using general dictionaries are likely to be imprecise and may produce underfit models that have a low level of predictive accuracy [35]. The business contexts in which authors develop their documents are also important. For example, the SEC staff have different foci in reviewing financial reports from firms in other industry sectors. These staff may have different writing preferences—Republican SEC staff often mention "unintended consequences of regulation," whereas Democratic SEC staff often mention "board diversity" [38]. Therefore, a domain-specific dictionary is important for textual analysis because it allows for a more accurate and relevant analysis of text within a specific field or subject area.

To our knowledge, at the time of writing, no dictionary specific to SEC reviews is available. A dictionary developed by author groups having an objective similar to ours is that of Bodnaruk et al. [22]. These authors focused on annual reports and identified words that can predict firm difficulties in raising money in the financial market. Their dictionary contains 184 words, 153 of which are not included in the Loughran and McDonald [17] dictionary; for example, "restraint," "prohibit," "unavailable," "mandate," and "dictate." However, their approach to dictionary development involves manual coding, which is likely to be labor intensive and prone to errors.

Thus, the commonly used sentiment-based dictionaries focus on

[3] The Multi-Perspective Question Answering Opinion Corpus contains news articles from a wide variety of news sources that are manually annotated for opinions and other private states (e.g., beliefs, emotions, sentiments, speculations). The National Research Council Canada Emotion Lexicon has a list of words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The annotations for the National Research Council Canada Emotion Lexicon were conducted manually by crowdsourcing. Quantitative Discourse Analysis Package is an R package designed to assist in quantitative discourse analysis.

[4] For example, pleasure, pain, arousal, overstated, political, interpersonal relations, and need.

[5] Diction has 31 wordlists, including ambivalence, tenacity, and inspiration. The lengths of the wordlists vary from 10 to 745 words, and no words are included in more than one list.

[6] Tone refers to the writer's overall attitude conveyed in a document. Loughran and McDonald [17] relied on their sentiment-based classification of words to measure the tone of firm annual reports.

[7] For example, in Google's annual report for the year 2010, management used the word "outstanding" in 19 sentences (e.g., "we had $3.0 billion of commercial paper outstanding"); this does not express a positive sentiment. The report can be found at https://www.sec.gov/Archives/edgar/data/1288776/000119312511032930/d10k.htm

Table 1
Framework for developing a dictionary specific to sentiment-devoid business documents.

Phase	Description
Phase 1: Defining the scope of the dictionary and identifying seed files	In this phase, researchers identify business documents from professional sources to be seed files for dictionary construction. Seed files are documents that researchers rely on to generate the wordlists of dictionaries. The relevant seed files should meet the following four criteria [40]: The seed files reveal information on business processes of interest or business documents and reports created by management or officials with domain knowledge. In prior studies not in the context of dictionary construction, researchers process management conference presentations and store traffic video files[25] to assess the quality of firms' accounting numbers. Researchers should not restrict sources of seed files to textual documents. They can consider voice, video, and email content, and convert this content to text and perform the procedures according to our dictionary-development framework. The literature does not provide guidelines for expected numbers of seed files. Researchers may consider using the "saturation" principle in qualitative research (i.e., when researchers see similar instances repeatedly, the researcher can have empirical confidence that seed files are saturated). The seed files should be of high quality. Given a focus on government and firm documents, researchers should seek relevant official sources. For example, they should consider audit reports by the Office of Inspector General of the US Department of Transportation or the *Guideline for Audit of IT Environment* by the European Court of Auditors in developing a dictionary for IT fraud; the *Health Insurance Portability and Accountability Act Privacy Rule* by the US Centers for Disease Control and Prevention or *General Data Protection Regulation* concerning health by the European Data Protection Supervisor for online healthcare privacy; and the documents by the US Office of the Comptroller of the Currency for bank IT security. In case researchers have a wide range of sources with different data formats, researchers may select those that are easier to process. Output of Phase 1: Seed files
Phase 2: Parsing text	Parsing seed files refers to using computer programs to convert unstructured content (e.g., raw texts in seed files) into structured content (e.g., words), which is more accessible for machine-learning algorithms. The bag-of-words model is a widely used method that simplifies the parsing process. The bag-of-words model is applied to seed files as follows: Exclude the noises in the textual data by converting all text into lowercase and excluding all nonword characters and punctuation. The model then outputs the textual data by tokenizing each sentence into words, generating vocabulary indices of words, and constructing the numerical feature vector that represents how frequently each word appears for each seed file. The data set for further processing to select words is an integrated matrix combined with the textual data for each seed file. The matrix consists of rows of seed file identifiers and columns of the frequency of each word that appears in any of the seed files. Output of Phase 2: Vectors
Phase 3: Ranking words	An identified typical adverse outcome in the domain can map the seed files likely to contain the words that express critical domain-specific information. The word-selection process for the final dictionary begins by excluding the most uninformative words. It keeps filtering out the words likely to contribute most to predicting the typical adverse outcome in the integrated matrix [35]. The word-selection process can be as follows: Exclude general stop words (e.g., "the") and sometimes numbers (e.g., "two"), names (e.g., "David"), and toponyms (e.g., "Boston") that are not the focus of the seed files' content and do not have incremental value for predictions. Exclude words with one or two letters used in the domain, such as "we" or "hi," which are not likely to be informative. Consider including words frequently appearing in the seed files published shortly before typical adverse outcomes occur because they can be candidate words. Select words by ensuring reliance on word importance (i.e., the extent to which a word contributes to prediction) beyond word frequency [62]. Assigning word importance should avoid cognitive biases and overcome knowledge limitations, which require automated methods. Output of Phase 3: Dictionary
Phase 4: Assessing dictionary quality	The quality of the dictionary can be assessed by applying the dictionary to examine some research questions in the domain. In addition, the dictionary should be compared with sentiment dictionaries. The assessment can be as follows: Develop the dictionary-based constructs for the research questions. Apply the dictionary and sentiment dictionaries to the documents for textual analysis in the literature of the domain and measure a construct based on each dictionary. Include the construct as the variable of interest in the empirical model to predict the dependent variable for the research question. The quality can be assessed using the regression results and comparisons between outcomes measuring the construct using the developed dictionary and other sentiment dictionaries, such as comparing the model's goodness of fit. Output of Phase 4: Supports quality assurance of dictionary

words that express positive or negative emotions of information creators. However, SEC documents contain very few sentiment words.[8] Given that SEC reviews shed light on information related to the quality of firms' financial systems and technology controls and cybersecurity,[9] a dictionary specific to SEC review documents can enable IS researchers to effectively infer meaning from textual content expressing the SEC's viewpoints and subsequently predict firms' performance from the SEC's perspective.

3. Framework for constructing domain-specific dictionaries

We followed Lee et al. [39] to develop a framework that consists of four phases: (1) defining the scope of the dictionary and identifying seed

[8] Ngai and Lee [63] reviewed 55 papers on the application of textual analysis of documents published in the government domain. By analyzing the sentiment expressed in documents generated by stakeholders, a government or government agency may be able to evaluate its practices. In contrast, although also focusing on a government agency (i.e., the SEC), our study explores the application of textual analysis on sentiment-devoid documents produced by the SEC, which is expected to be useful for analyzing the SEC's evaluations of firms.

[9] The SEC has issued guidelines and regulations on technology controls and cybersecurity, focusing on entities in the financial sector, including regulation systems compliance and integrity, Regulation S-P (privacy of consumer financial information), cybersecurity guidance, and regulation best interest.

files, (2) parsing text, (3) ranking words, and (4) evaluating the quality of the dictionary.[10] Table 1 provides an overview of each phase's goals and core activities. Fig. 1 provides a flowchart illustrating the dictionary-construction process.

3.1. Phase 1: defining the scope of the dictionary and identifying seed files

Phase 1 of constructing a domain-specific dictionary involves defining the scope of the research domain, including its extent and boundaries. It is then necessary to identify the key entities involved in communication within the domain, and these entities represent the primary actors or organizations that generate relevant data. For example, in SEC investigations, the key entities are SEC investigators and firm management. The theme of the dictionary we are creating will guide researchers' focus on the textual content generated by SEC investigators or firm management by refining the dictionary-construction process to suit the area of research. After the key entities involved are identified, researchers should outline the roles and duties of each entity, as well as the overall investigation process, to identify relevant and useful document sources for building the dictionary.

After identifying the domain's scope and key entities, researchers must determine the documents produced by entities pertinent to their study. The next step is to select relevant seed files from these documents. Seed files serve as sources of words for the dictionary and can be derived from various forms of media, such as text, audio, and video. However, no comprehensive guidelines exist in the literature for this selection process. Kearney and Liu [40] suggested adding seed files until they yield limited marginal information. Moreover, researchers should select seed files from primary sources (e.g., investigators) rather than from secondary sources (e.g., news articles) to ensure accuracy and richness of information [17,35]. This allows the dictionary to include the words that fit the language used by individuals in the domain. For example, in analyzing social media posts, researchers pay attention to intentional spelling "errors" that StockTwits users purposely employ, for example, misspelling "hold" as "hodl" to express keeping cryptocurrency, possibly including these words in their dictionaries. Researchers can find potential seed files in public and private documents prepared by information providers. By carefully selecting seed files and considering their sources, researchers can create a more robust foundation for their domain-specific dictionary construction.

3.2. Phase 2: parsing text

This phase aims to identify and prepare high-quality inputs for the machine-learning algorithms that subsequently rank words that will potentially be included in the domain-specific dictionary, akin to Lee et al.'s [39] step of identifying candidate words. This involves parsing raw text from seed files into quantitative data that reflect the linguistic properties of the files. However, raw texts are often unstructured and messy, making it difficult to identify linguistic properties directly. Consequently, using raw texts as inputs may lead to machine-learning algorithm underfitting [41].

There are various methods to extract linguistic properties from raw text and better preserve linguistic properties. These methods include label encoding, one-hot encoding, and bag-of-words modeling. Among these options, the bag-of-words model is widely adopted and easy to implement. It involves three major steps. In the first step, texts are standardized to minimize potential noise. In this standardization, all texts are converted to lowercase and punctuation is excluded, unless it has semantic value. In the second step, sentences are tokenized by space. Nonwords, such as document identification numbers, stop words, and short words, are generally excluded [35]. Stop words refers to the most commonly used words in a language that have little useful information.[11] In the third step, the remaining words are vectorized by assigning one dimension of a vector to one word and the word frequency to the dimension's value. Researchers may then select a cutoff for the count of seed files containing a word (i.e., document frequency) to exclude unimportant words in the domain. For example, Loughran and McDonald [17] use a cutoff of words used in at least 5 % of firm annual reports. However, this cutoff may omit rare but essential words. Alternatively, researchers may adopt a cutoff according to term frequency–inverse document frequency (TF-IDF), which highlights unique words relevant to a specific seed file's subject matter by considering the rarity of a word across the entire sample. However, TF-IDF might emphasize words too specific to a single seed file, limiting its applicability to predict across the entire sample. Researchers must make careful decisions aligning with their research focus when choosing a cutoff method. Nonetheless, using the cutoff as a primary filter can improve the efficiency of the subsequent ranking process in Phase 3.

3.3. Phase 3: ranking words

The primary goal of this phase is to identify and rank words critical for constructing a high-quality dictionary. Effective word selection is crucial to ensure comprehensive coverage of relevant terms and concepts. Traditional coding methods, such as manual annotation, often involve multiple annotators assigning importance scores to words, but these approaches can be inefficient and limited in domain-specific knowledge. Alternatively, machine-learning algorithms provide a more efficient and accurate means of ranking word importance by analyzing patterns in extensive text corpora and leveraging pretrained models [7]. Algorithms such as naive Bayes, neural networks, decision trees, maximum entropy, and support vector machines capture general language patterns and domain-specific nuances, resulting in more comprehensive and accurate dictionaries [23].

Neural networks, specifically, have gained attention for their ability to recognize patterns and extract relevant features from input data [42]. Unlike traditional approaches that rely on word frequency, neural networks assess a word's importance according to its contextual relevance [43]. This allows them to capture intricate relationships, even for infrequent words, resulting in a broader spectrum of domain-specific terms and concepts. In addition, neural networks can identify features and patterns not always apparent to human annotators, making them highly effective for dictionary development [7].

Further, Sheikh et al. [62] showed that integrating neural networks with the bag-of-words approach enhances text classification by identifying the most significant words for specific tasks. Converting texts into fixed-size vectors mitigates sensitivity to input length, improving classification accuracy. By blending bag-of-words with neural networks, researchers can capture lexical features and the subtle patterns identified by neural networks, leading to more accurate and domain-specific dictionaries.

In summary, machine-learning algorithms, particularly neural

[10] Our Phase 1 and Phase 2 correspond to the literature review step of Lee et al. [39], which aimed to comprehensively identify candidate words. Similarly to Lee et al. [[39], p. 1460], who employed an "empirical-to-conceptual" approach, we used machine learning to analyze a sufficient amount of data for categorizing these candidate words according to their importance in Phase 3. Finally, we performed multiple analyses to assess the quality of our dictionary, following their step of reviewing the categorization process. We detail our framework in the following subsections.

[11] Typical stop words include articles (e.g., "the," "a," and "an"), conjunctions (e.g., "and," "but," and "because"), prepositions (e.g., "at," "on," and "in"), and pronouns (e.g., "you," "your," and "yours"). Similarly, short words comprising one or two letters (e.g., "oh," "yo," or "hi") provide little information.

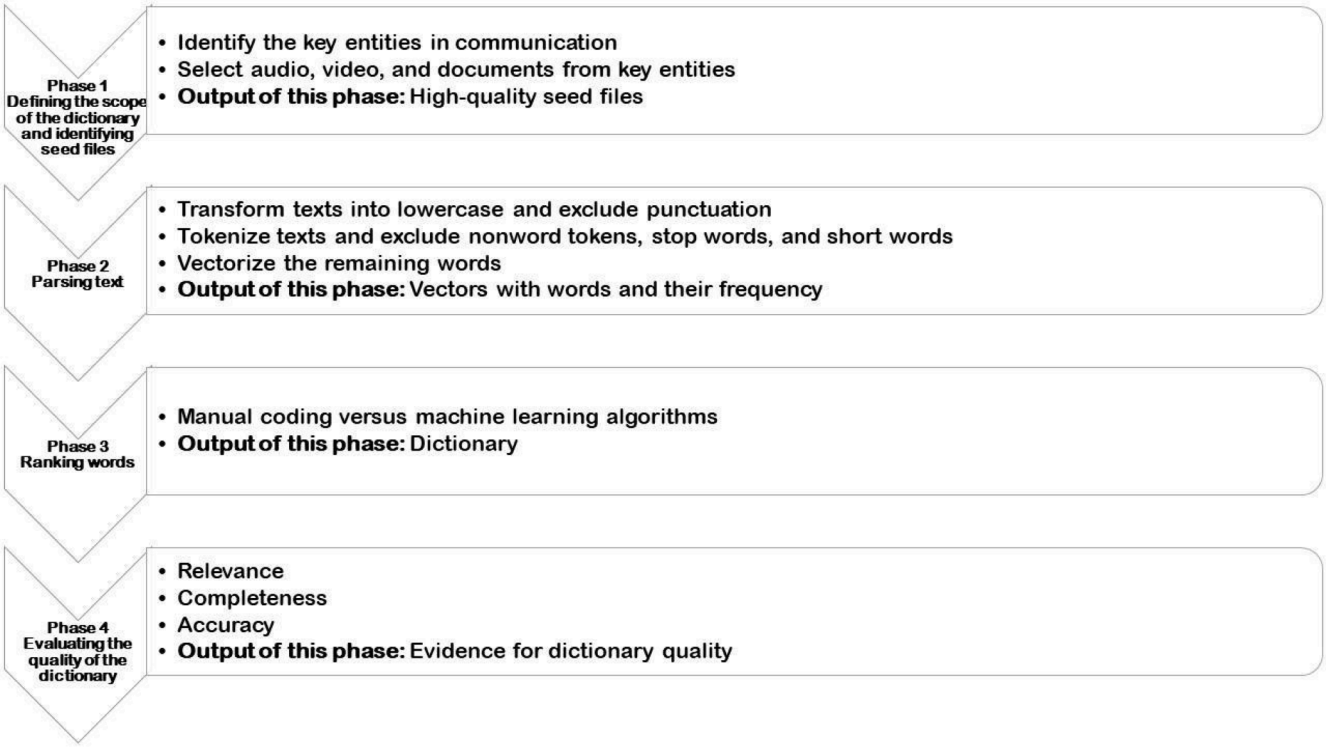


Fig. 1. Flowchart of dictionary-construction process.

networks, are powerful tools for ranking word importance because of their ability to recognize complex patterns and extract relevant features from text data.[12]

3.4. Phase 4: evaluating the quality of the dictionary

Phase 4 involves evaluating the quality of the dictionary. We incorporate the dimensions of high-quality information identified by Demoulin and Coussement [44] into this phase, assessing the dictionary using three key criteria from their framework: relevance, completeness, and accuracy.[13] First, relevance refers to how well the words in the dictionary align with the research domain and meet users' needs in understanding and applying domain-specific language and concepts [45]. Demoulin and Coussement [44] operationalized relevance by evaluating how effectively the output from machine-learning algorithms or textual analysis corresponds to users' research objectives. In the context of SEC investigations, a relevant dictionary should help researchers assess the severity of a firm's problems as perceived by the SEC [4]. Therefore, a high-quality dictionary must identify key aspects of the SEC's investigation that allow for the quantitative measurement of problem severity.

Second, completeness refers to how thoroughly the dictionary captures the researchers' specific interests within the textual data [46]. A comprehensive dictionary for SEC investigations must cover a broad spectrum of key terms used in SEC CLs to describe the investigation's scope and focus. Completeness is demonstrated when the dictionary adequately reflects the critical content found in the SEC's communications.

Third, accuracy measures the performance of a model by dividing the number of correctly predicted or classified instances (true positives and true negatives) by the total number of instances. For an SEC-specific dictionary, accuracy can be evaluated by how well it predicts investigation outcomes, such as identifying ITCWs. If the dictionary accurately predicts a high likelihood of ITCWs in cases in which they are later confirmed, and predicts a low likelihood where few ITCWs occur, it is considered highly accurate.

4. An illustration: developing a dictionary using SEC CLs

This example outlines a four-phase process for developing a dictionary tailored to the textual analysis of SEC investigations into firms' information quality, information system deficiencies, and financial discrepancies. We focused on capturing the SEC's perspective rather than firm responses. To achieve this, we processed SEC CLs to create a dictionary designed specifically for analyzing SEC investigations.

4.1. Phase 1: defining the scope of the dictionary and identifying seed files

In Phase 1, our primary goal is to gather seed files that will serve as the foundation for our dictionary. This starts by identifying the key entities involved in an SEC investigation, as illustrated in Fig. 2. SEC investigations can be initiated through multiple avenues, such as whistleblower tips, inquiries, or complaints and suspicions about a firm's accounting or business practices. Investigations may also arise from bankruptcy cases or other transaction-related documents.

[12] When it is necessary to update a dictionary, an incremental learning approach can be applied to neural network modeling. For instance, researchers can adapt the neural network's input layer by adding neurons for new words and expanding the first hidden layer to establish connections with these neurons. The weights of existing connections are initially kept constant while only the new connections are initialized. After this initial phase, gradual fine-tuning can be performed, where the existing weights are allowed to adjust with a very small learning rate, while the new connections are updated more rapidly. This helps preserve the model's previously learned knowledge while integrating new information. Finally, the dictionary can be updated based on the feature importance of the new words, as determined by the output of the fine-tuned neural network model.

[13] Demoulin and Coussement [44] proposed a framework to assess the quality of consumers' textual information when examining how information quality affects management's intention to use text-mining software to support decision-making.

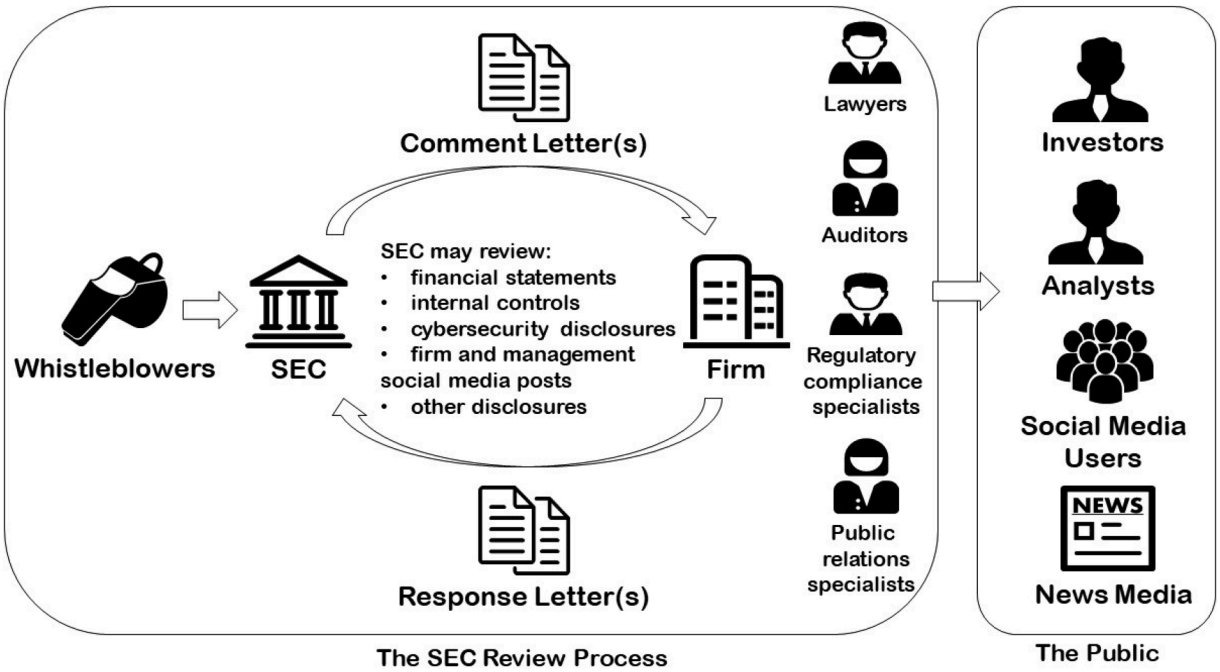


Fig. 2. Illustration of the SEC review process and key entities.

Once an investigation is initiated, the SEC sends a series of CLs to the firm under investigation. The first, known as the "master letter," outlines the SEC's concerns. The firm must respond within 10 business days with a "response letter." The SEC may then issue follow-up CLs in accordance with the firm's reply. If the SEC is satisfied with the response about information system deficiencies, a CL concluding the review is sent. If not, communication continues until the issues are resolved.

During the investigation, external entities, such as investors and social media users, may weigh in with their perspectives. By analyzing these SEC CLs and related materials, we can identify further seed files to form the basis of our dictionary, helping us define relevant keywords and concepts central to the investigation.

Our goal was to create a dictionary that offers researchers deeper insights into the SEC's perspective on firm-related issues. We focused our analysis on the SEC's master letters, which are legal documents written in a precise and professional manner.[14] Master letters were chosen as seed files for several reasons. First, they are authored by SEC staff, who play a crucial role in enforcing regulations on control deficiencies and inadequate disclosures [13]. Their authority and expertise ensure that master letters are of exceptional quality. Second, these letters comprehensively address a range of disclosure issues [13]. Third, unlike follow-up letters, master letters provide the most thorough assessment of a firm's IS [13].

By examining SEC reviews conducted between 2005 and 2019, we extracted 78,926 master letters related to various firm outputs, including annual and quarterly reports, proxy statements, and registration statements. However, prior studies have focused primarily on master letters related to annual and quarterly reports because these documents offer a comprehensive view of the information that firms' systems should generate [4,14]. These reports also provide insights into firms' financial performance, risk assessments regarding technology and cybersecurity, and discussions on how technology is used to drive innovation and improve efficiency [15,16]. Therefore, we narrowed our focus to 26,494 master letters related to annual or quarterly reports, containing 64,087 unique words.

Table 2
CL sample selection and candidate word pool construction.

	CLs (1)	Unique words (2)
Filtering the seed files		
All CLs from 2005 to 2019	150,432	188,595
After excluding nonmaster letters	78,926	138,996
After excluding CLs unrelated to annual or quarterly reports	26,494	64,087
Filtering the words for the pool of candidate words		
After excluding words used in less than 1 % of CLs	–	2,660
After excluding short words and stop words	–	1,868
Final dictionary with above-median feature importance determined by the neural network model in predicting *RV_RES*	–	934

4.2. Phase 2: parsing text (Using bag-of-words for candidate words)

To analyze the seed files, we employed a bag-of-words model using an R program that processed each of the 26,494 CLs, converting them into individual vectors. Each vector represented a specific CL; distinct words in rows and columns indicated the frequency of each word's occurrence. Following the methodology of Loughran and McDonald [17], we identified candidate words for the dictionary. Table 2 shows the number of CLs and identifies words at various stages of the dictionary-construction process. Initially, we constructed a pool of candidate words and applied a 1 % cutoff to exclude words that rarely appeared in the CLs, resulting in 2,660 unique words. We then removed stop words to reduce noise, refining the pool to 1,868 unique words. Finally, we converted the frequency data of these 1,868 words into a document-term matrix, in which each row represented a unique CL and each column indicated the frequency of a specific word within that CL.[15]

[14] The nature of SEC investigations ensures a uniform objective among document authors, distinguishing the documents produced by the SEC from crowdsourced documents that are considerably influenced by the different motivations of their authors [64].

[15] For example, we defined revisei,t as the count of the word "revise" in a master letter, scaled by the total word count of the CL.

4.3. Phase 3: ranking words (for predicting restatements)

In Phase 3, we identified relevant words from the constructed matrix to build the dictionary, ranking them according to their ability to predict the likelihood of financial restatements. We chose financial restatements as our predictive outcome because they occur when management corrects errors identified by the SEC, signifying the severity of the firm's problems [4]. This aligns with Ryans' argument that financial restatements represent an adverse outcome closely related to the themes of SEC investigations.

To rank words, we developed a prediction model using neural network analyses. Neural networks are highly effective for sentiment and textual analysis because of their ability to automatically extract features from large data sets [47]. This eliminates the need for labor-intensive manual coding and allows for the identification of important words and phrases. Neural networks also excel at handling extensive vocabularies, prioritizing words according to their significance for specific tasks [48]. Advanced models, such as recurrent neural networks and transformers, capture contextual nuances, making them particularly useful for sentiment analysis [49]. Further, their capacity for continuous learning ensures that the dictionary remains up-to-date and relevant [50].

Neural networks often outperform traditional methods, such as ordinal logistic regression, because of their lower and more stable misclassification rates [42]. They have been successfully applied to predict emotions, attitudes, and decision-making behaviors [51]. For instance, Ghiassi et al. [52] used neural networks to create a sentiment dictionary from Twitter data, and Kraus and Feuerriegel [53] and Mahmoudi et al. [54] identified predictive words from firm disclosures and investor microblogs using similar techniques.

In our neural network model, the dependent variable was *RV_RES*, which equaled 1 if management announced a financial restatement between the commencement date of the SEC review and the completion date of the review, and 0 otherwise. Specifically, we considered an SEC investigation to commence at the issuance date of the master letter and end at the dissemination date of the documents on EDGAR. Financial restatements during this period are likely triggered by the SEC investigation and indicate severe problems [4]. In our tests, the treatment group included the 1,683 master letters that $RV\_RES = 1$. We selected the 1,683 master letters that $RV\_RES = 0$ to be included in our control group. The two groups together comprised a total of 3,366 master letters.

We used the frequency of 1,868 words as the covariates in our neural network model to predict whether financial restatements would occur during each SEC review.[16] We randomly selected 70 % of our final sample as the training sample and used 30 % of that as the testing sample. The neural network model included two hidden layers and had 20 neurons in each layer.[17]

Table 3 presents the performance metrics of our neural network model. One key metric is the area under the curve (AUC) of the receiver operating characteristics, or AUC score. Our model achieved an AUC of 0.937, indicating that it accurately prioritizes nearly 94 % of cases according to the likelihood of requiring financial restatements. This robust performance significantly outperforms random assignment (AUC = 0.5), demonstrating the model's effectiveness in distinguishing between cases with and without financial restatements.

Table 3
Neural network modeling performance.

Area under the curve	0.937
Training phase	
Sample	2,347
Accuracy	0.911
Precision	0.918
Recall	0.889
F-score	0.903
Testing phase	
Sample	1,019
Accuracy	0.857
Precision	0.860
Recall	0.856
F-score	0.858

During the training phase, the model demonstrated strong classification accuracy at 0.911. Its precision of 0.918 indicates that when the model predicts a financial restatement, it is accurate 91.8 % of the time. In addition, the recall score of 0.889 shows that the model correctly identifies 88.9 % of actual financial restatement cases. The F1 score of 0.903, a harmonic mean of precision and recall, confirms that the model strikes a good balance between these two metrics, achieving an overall performance of 90.3 % . In the testing phase, the model maintained a high performance; all metrics consistently exceeded 0.850. Overall, our neural network model is highly effective at identifying textual cues in SEC CLs that signal severe deficiencies in firms' systems.

Finally, we ranked the words by their feature importance as determined by the neural network model. From this ranking, we selected the 934 words that had importance values above the median. These words, deemed highly influential, were included in our final dictionary because they serve as strong indicators of severe system deficiencies within a firm.

4.4. Phase 4: evaluating the quality of the dictionary

4.4.1. Description of evaluation process

We assessed the quality of our constructed dictionary by using it for textual analyses of the SEC's CLs and evaluating its predictive capability in determining a focal firm's likelihood of ITCWs, IT audit fees, and cyber risks. In addition, we conducted comparative analyses using five other dictionaries—the Loughran and McDonald [17] dictionary, the Harvard General Inquirer dictionary, the CFI financial dictionary, the Investopedia financial dictionary,[18] and the Henry [21] dictionary—to benchmark the performance of our dictionary against these established resources.

Our first focus was identifying ITCWs among various firm problems for two key reasons. First, ITCWs are of significant interest to IS researchers, given the critical role of system deficiencies and security in this field. Predicting ITCWs is essential because inadequate IT controls can severely affect the quality of information produced by management IS (C. [3]). Second, identifying and remediating ITCWs is a fundamental responsibility of the SEC. Once ITCWs are identified, it is likely that SEC-guided reviews, along with management and auditor assessments, will uncover additional deficiencies within the firm's IS.

Our second focus was evaluating IT audit fees, driven by the growing reliance on IT services for key business operations and recent legislative and professional requirements regulating IT audits [55]. The fees associated with IT audits reflect auditors' assessments of the risks involved in ensuring firms' compliance with IT controls [56]. We expect that

[16] We normalize each covariate before training.

[17] The existing literature does not provide a definitive rule for determining the optimal number of neurons in the hidden layers of a neural network. Srivastava et al. [41] proposed an approach called "dropout," which randomly reduces the number of neurons to mitigate the risk of overfitting caused by an excess of neurons. By implementing this dropout technique on the hidden layers, we discovered that the model's performance was optimized by having two hidden layers, each containing 20 neurons. This optimal architecture not only conforms with the theoretical underpinning of the dropout method but also aligns with the architecture automatically generated by our neural network program.

[18] We thank the anonymous reviewer for the valuable suggestion to use the CFI financial dictionary and the Investopedia financial dictionary for benchmarking in our study.

auditors' risk assessments—and, consequently, audit fees—are influenced by SEC reviews, particularly if a firm is deemed high risk.

Our third focus was on firms' cyber risks, an area that has received increasing attention from the SEC [15]. Firms are often reluctant to disclose cybersecurity breaches because of concerns about financial and reputational damage. The SEC may highlight concerns regarding a firm's cybersecurity controls during its reviews [15]. We posited that SEC CLs may contain textual clues that can help predict a firm's exposure to cyber risks.

4.4.2. Analysis models

We adopted the methodology of Ryans [4] to evaluate the accuracy of our dictionary. Specifically, we used our dictionary and the five benchmarking dictionaries to develop constructs that measured the severity of SEC CLs.[19] We calculated the frequency of matched words and normalized this count by dividing it by the total word count of the letter, denoted as *MH*. An increase in the value of *MH* was a proxy for the elevated severity of information system deficiencies perceived by SEC staff within the firm.[20] *MH* was expected to exhibit a positive correlation with the likelihood of ITCWs occurring in the following year.

Similarly, we applied the five other dictionaries to the SEC's master letters. We measured *LM* as the count of the words included in the Loughran and McDonald [17] dictionary's negative sentiment wordlist, scaled by the master letter's total word count.[21] *HarvardGI* denotes the proportion of negative sentiment words from the Harvard General Inquirer dictionary in the master letter. In addition, *CFI* and *Investopedia* were measured in the same way, using the CFI financial dictionary and the Investopedia financial dictionary, respectively. Finally, we defined *Henry* as the frequency of negative sentiment words from the Henry [21] dictionary in the master letter. If the words from the five other dictionaries, particularly those expressing negative sentiment, were effectively applicable to the context of SEC investigations, the five variables for these dictionaries may also have a positive association with the outcomes in the following year. We referenced Asthana et al. [57], W. Y. Li et al. [10], and Mazza and Azzali [56] to include the variables for the factors that are commonly controlled in the literature of ITCW, IT audit fees, and cyber risks. Model (1) is as follows:

$$\begin{aligned} Dependent_{i,t+1} = {} & \beta_0 + \beta_1 Independent_{i,t} + \beta_2 ICW_{i,t+1} + \beta_3 LogAsset_{i,t+1} \\ & + \beta_4 Leverage_{i,t+1} \\ & + \beta_5 Current_{i,t+1} + \beta_6 ROA_{i,t+1} + \beta_7 Cash_{i,t+1} + \beta_8 Loss_{i,t+1} \\ & + \beta_9 Big4_{i,t+1} + \beta_{10} Restate_{i,t+1} + IndustryFE + YearFE + \varepsilon_{i,t} \end{aligned} \quad (1)$$

We provide comprehensive variable definitions in Appendix A. $Dependent_{i,t+1}$ denotes each of the dependent variables we examined when evaluating the quality of our dictionary, including *ITCW*, *IT_Audit_Fees*, and *CyberRisk*. *ITCW* serves as an indicator variable. It was assigned a value of 1 if the firm identified an *ITCW* one year after the completion of the SEC review, and 0 otherwise. *IT_Audit_Fees* was measured as the IT audit fees paid to auditors in one year following the SEC review completion. To measure *CyberRisk*, we used the model's value predicted by Asthana et al. [57] in one year after the SEC review was completed. $Independent_{i,t}$ denotes each of our independent variables of interest, which include the variables for the proportion of words from each dictionary in the master letter: *MH, LM, HarvardGI, CFI, Investopedia*, and *Henry*.

Our model included the control variable *ICW*, which denotes the number of other types of control weaknesses identified within a year of the completion of the SEC review. The effectiveness of general internal controls significantly affects the weaknesses and risks present in a firm's information system. Auditors evaluate the effectiveness of general internal controls to establish the fees for their related services.

Further, we included control variables to address firms' financial constraints that could impede the effectiveness of controls and risk management. These variables include firm size (*LogAsset*), default risk (*Leverage*), liquidity position (*Current*), operational performance (*ROA* and *Loss*), and cash holding (*Cash*). Larger firms typically possess more substantial resources, enhancing their ability to ensure robust control effectiveness and risk management. However, auditors tend to demand higher fees from larger firms because of the complexity of the services required. This complexity stems from the intricate organizational structures and diverse operational characteristics of larger firms. Thus, we measured *LogAsset* as the natural logarithm of total assets to control for firm size. *Leverage*, measured as total liabilities scaled by total assets, was a representation of default risk. A high *Leverage* value can suggest notable financial pressure from creditors and potential diversion of resources from IT controls and risk management. We measured *Current* as current assets, and *Cash* as cash and cash equivalents scaled by total assets. For liquidity, a low value of *Current* and *Cash* could indicate liquidity constraints, which may adversely affect a firm's ability to fulfill short-term obligations. Concerning operational performance, a firm's profitability often leads to a larger IT budget. However, this can render the firm more susceptible to cyberattacks. To capture this, we measured *ROA* as net income scaled by total assets, and *Loss* as an indicator variable that equaled 1 if the firm reported negative net income, and 0 otherwise.

In addition, we considered the variable *Big4*, an indicator variable set to 1 if a Big 4 auditing firm conducted the firm's audit in the year following the completion of an SEC review, and 0 otherwise. Given their extensive IT expertise, Big 4 auditors are anticipated to provide superior services to assist firms in identifying ITCWs and managing cyber risks. This is even though their services generally come at a higher price. Finally, we controlled for potential information deficiencies within firms using *Restate*, an indicator variable set to 1 if the firm restated its publicly disclosed information one year after the completion of the SEC review, and 0 otherwise.

4.4.3. Sample and descriptive statistics

Our sample comprised 19,791 SEC reviews conducted between 2005 and 2021, excluding any reviews missing data for variables included in our models. Descriptive statistics for dictionary-related variables are presented in Panel A of Table 4. The mean value of *MH* was 16.560, suggesting that, on average, 16.56 % of the content in an SEC master letter pertains to the words from our dictionary. *MH* exhibited the highest standard deviation compared with the other five dictionary-related variables, indicating significant variability in how our dictionary's words were represented across SEC master letters. This suggested that our dictionary has the potential to capture the diverse content present in these letters.

Although the Loughran and McDonald [17] dictionary and the Harvard General Inquirer dictionary contain more words than ours, 46.96 % and 65.09 % of their words, respectively, are never used in SEC CLs. This suggests that approximately half of the words included in these dictionaries can be considered irrelevant to the domain of business and

[19] Ryans [4] used machine-learning algorithms to classify the SEC's master letters into "important" and "innocuous" categories in predicting the occurrence of financial restatements during SEC reviews. Ryans [4] assigned a binary variable to master letters deemed important, taking a value of 1, and 0 otherwise. The binary variable was inputted into a Probit regression model to predict the occurrence of financial restatements in the following year. Similarly, we used our dictionary to predict the outcomes in the year following SEC reviews.

[20] To evaluate the completeness of our dictionary, we included an additional 5%, 15%, or 25% of candidate words. These words were ranked immediately after those initially selected for our dictionary. However, this inclusion of additional words did not lead to any significant increase in the pseudo-R-squared values of our analyses. This outcome suggested that the marginal contribution to explaining the variance in related variables by these extra words is limited, thereby indicating the completeness of our dictionary.

[21] In the context of SEC investigations, we focused on words that had a negative sentiment.

Table 4
Summary statistics and correlation matrix.

Panel A. Dictionary-related variables (19,791 Observations)

	Word number	Zero-occurrence word percentage	Mean	S.D.	P1	Median	P99
MH	934	0.00	16.560	5.556	3.883	17.032	27.697
LM	2,355	46.96	2.351	1.217	0.000	2.116	6.230
HarvardGI	2,005	65.09	1.856	1.124	0.000	1.744	5.226
CFI	114	15.79	3.520	2.213	0.000	3.226	9.756
Investopedia	5,564	39.50	56.302	3.423	48.106	56.419	63.953
Henry	85	20.00	0.277	0.427	0.000	0.000	1.872

Panel B. Dependent variable and control variables

	Observations	Mean	S.D.	P1	Median	P99
ITCW	19,791	0.026	0.161	0.000	0.000	1.000
IT_Audit_Fees	19,791	0.742	1.795	0.000	0.148	12.900
CyberRisk	17,170	12.719	30.353	0.012	3.065	147.011
ICW	19,791	0.398	1.402	0.000	0.000	7.000
LogAsset	19,791	7.095	2.544	−0.835	7.353	12.637
Leverage	19,791	0.705	0.767	0.064	0.605	6.778
Current	19,791	1.430	3.842	0.000	0.185	27.730
ROA	19,791	−0.126	0.745	−6.098	0.022	0.326
Cash	19,791	0.161	0.194	0.000	0.084	0.892
Loss	19,791	0.313	0.464	0.000	0.000	1.000
Big4	19,791	0.730	0.444	0.000	1.000	1.000
Restate	19,791	0.094	0.291	0.000	0.000	1.000

Notes: This table presents the summary statistics for all variables included in our models. *ITCW* is an indicator that equals 1 if the firm reports any ITCW in the year following the SEC investigation, and 0 otherwise. *IT_Audit_Fees* is measured as the value of the fees paid to auditors that are not related to financial statement audits in the year following the SEC investigation. *CyberRisk* is measured as the predicted value of the model of Asthana et al. [57] in the year following the SEC investigation. *MH* is measured as the count of words from our constructed dictionary that are found in the master letter, divided by the letter's total word count. *LM* is measured as the count of negative sentiment words from the Loughran and McDonald [17] dictionary found in the master letter, divided by the letter's total word count. *HarvardGI* is measured as the count of negative sentiment words from the Harvard General Inquirer dictionary that are found in the master letter, divided by the letter's total word count. *CFI* is measured as the count of words from the CFI financial dictionary found in the master letter, divided by the letter's total word count. *Investopedia* is measured as the count of words from the Investopedia financial dictionary that are found in the master letter, divided by the letter's total word count. *Henry* is measured as the count of words from the Henry [21] dictionary found in the master letter, divided by the letter's total word count. *ICW* is measured as the number of internal control weaknesses unrelated to IT that the firm reports in the year following an SEC investigation. *LogAsset* is measured as the natural logarithm of total assets. *Leverage* is measured as the ratio of total liabilities to total assets. *Current* is measured as the current assets. *ROA* is measured as the ratio of net income to total assets. *Cash* is measured as the ratio of cash to total assets. *Loss* is an indicator that equals 1 if the firm reports negative net income, and 0 otherwise. *Big4* is an indicator that equals 1 if the firm is audited by a Big 4 auditor, and 0 otherwise. *Restate* is an indicator that equals 1 if the firm makes any financial restatements, and 0 otherwise. The detailed definition of variables can be found in Appendix A.

governance, which means that our dictionary has better relevance. In addition, the Henry [21] dictionary and the Investopedia financial dictionary had the lowest (0.277) and highest (56.302) frequency mean among the six dictionaries. This might suggest the Henry [21] dictionary's limited relevance or the Investopedia financial dictionary's loss of specificity in the domain of business and governance.

Panel B of Table 4 provides the descriptive statistics for the dependent and control variables. In the year following SEC investigations, identified ITCWs were present in 2.60 % of cases. In addition, 0.398 other types of internal control weaknesses were identified on average. The subsequent average IT audit fees amounted to USD 0.742 million, and 73.00 % of these fees were attributed to services provided by Big 4 auditors.[22]

4.4.4. Regression results

4.4.4.1. Comparison between dictionaries in predicting ITCW. Table 5 presents the comparative performances of our dictionary and the Loughran and McDonald [17] dictionary in predicting future ITCW. Column (1) presents the results of a logistic regression estimating Model (1) using *MH* as the independent variable. The results in this column suggest a positive association between the proportion of words from our dictionary in a master letter and the likelihood of identifying an ITCW in the following year (coef. = 0.037, $p < 0.05$). Specifically, when this proportion increases by one unit, the likelihood of an ITCW being identified in the following year increases by 3.70 % .[23] This indicated that our dictionary captures severe problems in a firm's IS identified during an SEC review and predicts that the firm's management and auditors will be motivated to review IT controls and disclose any identified weaknesses. For this model, the log pseudo-likelihood is −1, 491.2845 and the pseudo-*R*-squared is 38.03 % .

In addition, we fail to find any evidence of statistically significant effects of the proportion of words from the five other dictionaries on the likelihood of identifying an ITCW in the following year. In Columns (2) to (6) of Table 5, none of the models for these five dictionaries demonstrate a value of the pseudo-*R*-squared higher than 38.03 % . For example, in Column (2), the value of the pseudo-*R*-squared is 37.86 % , suggesting that when the dictionaries are applied to SEC CLs, the Loughran and McDonald [17] dictionary underperforms our dictionary by 0.45 % in explaining ITCW likelihood. Overall, our dictionary performed the best in predicting firms' future ITCW likelihood.

4.4.4.2. Comparison between dictionaries in predicting IT audit fees. Table 6 presents a comparative analysis of the performance of our dictionary against those of the other five dictionaries in predicting future IT audit fees. Column (1) presents the results from an ordinary least squares regression, estimating Model (2) using *MH* as the independent variable. The results indicated a positive relationship between the proportion of words from our dictionary in a master letter and the IT audit fees for the

[22] In an untabulated analysis of the variance inflation factor, we found that none of the variance inflation factor values exceeded 4. This suggested that our model is unlikely to be significantly affected by multicollinearity concerns.

[23] The odds ratio of MH was 1.037.

Table 5
Comparison between the developed dictionary and five other dictionaries in predicting ITCW in the following year.

Dependent variable: *ITCW*						
	(1)	(2)	(3)	(4)	(5)	(6)
MH	0.037**					
	(0.019)					
LM		0.027				
		(0.068)				
HarvardGI			0.049			
			(0.071)			
CFI				0.045		
				(0.034)		
Investopedia					0.029	
					(0.026)	
Henry						0.229
						(0.154)
ICW	0.664***	0.666***	0.667***	0.665***	0.665***	0.666***
	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
LogAsset	0.100	0.093	0.091	0.086	0.092	0.091
	(0.065)	(0.064)	(0.066)	(0.066)	(0.065)	(0.065)
Leverage	0.018	0.013	0.012	0.015	0.013	0.013
	(0.125)	(0.125)	(0.125)	(0.126)	(0.125)	(0.126)
Current	−0.156**	−0.156**	−0.156**	−0.154**	−0.153**	−0.157**
	(0.073)	(0.074)	(0.074)	(0.074)	(0.073)	(0.074)
ROA	0.036	0.030	0.029	0.030	0.033	0.029
	(0.139)	(0.138)	(0.138)	(0.138)	(0.139)	(0.139)
Cash	−0.581	−0.600	−0.604	−0.589	−0.598	−0.603
	(0.486)	(0.484)	(0.484)	(0.484)	(0.486)	(0.484)
Loss	0.024	0.033	0.033	0.035	0.039	0.034
	(0.200)	(0.201)	(0.200)	(0.200)	(0.199)	(0.200)
Big4	−0.572**	−0.601**	−0.603**	−0.601**	−0.599**	−0.595**
	(0.247)	(0.244)	(0.246)	(0.245)	(0.244)	(0.244)
Restate	0.350	0.353	0.357	0.358	0.356	0.351
	(0.240)	(0.239)	(0.240)	(0.240)	(0.239)	(0.240)
Year fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Cluster by firm	Yes	Yes	Yes	Yes	Yes	Yes
Observations	19,791	19,791	19,791	19,791	19,791	19,791
Pseudo-*R*-squared	0.3803	0.3786	0.3787	0.3792	0.3792	0.3792
Log pseudo-likelihood	−1,491.2845	−1,495.3821	−1,495.0460	−1,493.8763	−1,494.0081	−1,493.9007

Notes: This table presents the results of a comparison between our dictionary and five other dictionaries in predicting a firm's ITCWs in the following year. The detailed definition of variables can be found in Appendix A. The standard errors are clustered at the firm level and reported in parentheses. *, **, and *** denote significance at the 10 % , 5 % , and 1 % levels, respectively.

subsequent year (coef.= 0.016, $p < 0.01$). This finding not only was statistically significant but also economically significant; a one-unit increase in this proportion corresponded to an additional fee of USD 0.016 million for IT audits. For this model, the *R*-squared value was 33.79 % , which is higher than those in Columns (2) to (6), indicating that our dictionary accounts for a larger proportion of the variation in IT audit fees than any of the other five dictionaries. Columns (2) and (3) present the models for the Loughran and McDonald [17] dictionary and the Harvard General Inquirer dictionary, respectively, both revealing the lowest *R*-squared values at 33.67 % . This highlighted that our dictionary outperforms these dictionaries in predicting future IT audit fees by up to 0.36 % .

Further, the coefficients on the proportion of words from the CFI financial dictionary, the Investopedia financial dictionary, and the Henry [21] dictionary were statistically significant, suggesting that these dictionaries may also have some predictive power for future IT audit fees. However, in untabulated results, *MH* has the largest standardized coefficient among the six variables representing the proportion of words from each dictionary. Specifically, the standardized coefficient for *MH* was more than double those for *CFI* and *Investopedia*. Taken together with *R*-squared values, the standardized coefficients demonstrated that our dictionary has the highest predictive power among the six dictionaries being investigated in predicting subsequent IT audit fees.

4.4.4.3. Comparison between dictionaries in predicting cyber risks. Table 7 compares the performance of our dictionary with those of five others in predicting future cyber risks. Column (1) shows the results of an ordinary least squares regression, estimating Model (3) using *MH* as the independent variable. The findings revealed a significant and positive relationship between the proportion of words from our dictionary in a master letter and the firm's cyber risks in the following year (coef.= 0.147, $p < 0.01$). Specifically, a one-unit increase in this proportion leads to a 1.16 % rise in mean cyber risks for the next year, indicating that our dictionary effectively identifies potential issues in a firm's cybersecurity systems during SEC reviews. Columns (2) to (6) show statistically insignificant coefficients for the other five dictionaries. The *R*-squared value is highest in Column (1).

In untabulated results, when the proportions of words from all six dictionaries were estimated simultaneously, *MH* remained the only statistically significant variable and had the highest standardized coefficient value among the six. These results demonstrated the superior performance of our dictionary in explaining cyber risks compared with those of the five other dictionaries.[24]

[24] The untabulated results show that using the information security dictionary developed by Gordon et al. [59] as an additional comparative dictionary does not affect our findings.

[25] In January 2020, Muddy Water announced an anonymous report titled "Luckin Coffee: Fraud + Fundamentally Broken Business," alleging that Luckin Coffee committed fraud. The report shows an analysis based on 11,200 hours of videos of surveillance and traffic in 620 stores. Details can be found at https://www.qsrmagazine.com/food/beverage/luckin-coffee-faces-fraud-allegations-anonymous-report/. We viewed the website on August 30, 2022.

Table 6
Comparison between the developed dictionary and five other dictionaries in predicting IT audit fees in the following year.

Dependent variable: *IT_Audit_Fees*						
	(1)	(2)	(3)	(4)	(5)	(6)
MH	0.016***					
	(0.004)					
LM		0.009				
		(0.012)				
HarvardGI			0.008			
			(0.015)			
CFI				0.026***		
				(0.007)		
Investopedia					0.017***	
					(0.005)	
Henry						0.101***
						(0.034)
ICW	0.010	0.013	0.013	0.012	0.012	0.012
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
LogAsset	0.335***	0.332***	0.332***	0.329***	0.330***	0.331***
	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)
Leverage	0.102***	0.103***	0.103***	0.104***	0.102***	0.102***
	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)
Current	0.138***	0.137***	0.137***	0.138***	0.139***	0.137***
	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)
ROA	−0.232***	−0.232***	−0.232***	−0.234***	−0.229***	−0.233***
	(0.036)	(0.036)	(0.036)	(0.037)	(0.036)	(0.037)
Cash	0.380**	0.373**	0.372**	0.386**	0.377**	0.369**
	(0.182)	(0.181)	(0.181)	(0.182)	(0.181)	(0.181)
Loss	0.054*	0.057*	0.057*	0.055*	0.058*	0.055*
	(0.029)	(0.030)	(0.030)	(0.029)	(0.030)	(0.030)
Big4	−0.171***	−0.184***	−0.184***	−0.182***	−0.180***	−0.181***
	(0.063)	(0.063)	(0.064)	(0.063)	(0.063)	(0.063)
Restate	−0.101**	−0.098**	−0.098**	−0.099**	−0.098**	−0.099**
	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)
Year fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Cluster by firm	Yes	Yes	Yes	Yes	Yes	Yes
Observations	19,791	19,791	19,791	19,791	19,791	19,791
R-squared	0.3379	0.3367	0.3367	0.3376	0.3376	0.3372

Notes: This table presents the results of a comparison between our dictionary and five other dictionaries in predicting a firm's IT audit fees in the following year. The detailed definition of variables can be found in Appendix A. The standard errors are clustered at the firm level and reported in parentheses. *, **, and *** denote significance at the 10 % , 5 % , and 1 % levels, respectively.

4.4.5. Robustness check: the weight of words

In Section 4.4.4, we followed Loughran and McDonald [17] by using our dictionary to develop a construct based on word frequency. In this section, we further test the robustness of our findings by creating alternative constructs that incorporate word weighting. First, we implemented a TF-IDF weighted construct. For each word in our dictionary, term frequency represented its proportion within an SEC master letter while document frequency reflected the number of master letters in the sample that contained the word. We then calculated the inverse document frequency as the natural logarithm of the total number of SEC master letters divided by the document frequency plus one. *MH_TFIDF* was subsequently defined as the sum of the term frequency multiplied by the inverse document frequency for all the words from our dictionary in an SEC master letter. Second, we introduced another construct, weighted by the feature importance as determined by our neural network model. The count of each word from our dictionary in an SEC master letter was multiplied by its normalized feature importance. For a given SEC master letter, the value of *MH_Importance* was computed as the sum of the word count, weighted by importance, and multiplied by the inverse document frequency.

We then substituted *MH* with *MH_TFIDF* and *MH_Importance* to estimate subsequent ITCW, IT audit fees, and cyber risks. The results of this robustness check are presented in Table 8. Across all three outcomes, the results for *MH_TFIDF* and *MH_Importance* were consistent with our findings for *MH*, as detailed in Tables 5–7. The models in Columns (1) to (4) exhibit higher pseudo-*R*-squared and *R*-squared values than those using *MH*. This supported Loughran and McDonald's [17] suggestion that an effective weighting procedure "raises the significance of the word lists by improving the signal-to-noise in the lists" (p. 57). Further, all three models using *MH_Importance* demonstrated higher pseudo-*R*-squared and *R*-squared values than those employing *MH_TFIDF*. This highlighted the potential of our methodology to enable researchers to develop high-quality, dictionary-based constructs through the effective weighting of words by their feature importance.

5. Discussion

5.1. Research summary and contributions

To date, few dictionaries have been developed specifically for business communication. One exception is that of Loughran and McDonald [17], which focuses on sentiment words in annual reports to predict financial performance. However, sentiment-analysis tools may not effectively distinguish between sentiment-devoid language and language that conveys concerns, doubts, or criticism about compliance practices, cybersecurity risks, or other technology-related issues. As a result, relying solely on sentiment-based tools may lead to incomplete or inaccurate analyses, limiting the ability of IS researchers to examine IT-related documents from firms and governments and potentially compromising the validity of their findings.

We extended the traditional dictionary-construction framework, which focuses primarily on sentiment-based dictionaries and pays less attention to the application context. Drawing from the literature, we proposed a four-phase framework that emphasizes the importance of context and business communication. In Phase 1, we defined the application domain and identified high-quality seed files. Phase 2

Table 7
Comparison between the developed dictionary and five other dictionaries in predicting cyber risks in the following year.

Dependent variable: *CyberRisk*						
	(1)	(2)	(3)	(4)	(5)	(6)
MH	0.147***					
	(0.048)					
LM		−0.120				
		(0.201)				
HarvardGI			0.149			
			(0.237)			
CFI				0.149		
				(0.099)		
Investopedia					0.004	
					(0.078)	
Henry						−0.603
						(0.444)
ICW						
LogAsset	0.087	0.111	0.110	0.106	0.11	0.115
	(0.130)	(0.129)	(0.130)	(0.130)	(0.130)	(0.129)
Leverage	3.202***	3.190***	3.176***	3.162***	3.182***	3.194***
	(0.332)	(0.329)	(0.330)	(0.330)	(0.331)	(0.331)
Current	2.040***	2.054***	2.041***	2.051***	2.047***	2.053***
	(0.455)	(0.454)	(0.454)	(0.456)	(0.455)	(0.455)
ROA	2.924***	2.922***	2.923***	2.926***	2.923***	2.923***
	(0.352)	(0.352)	(0.353)	(0.353)	(0.353)	(0.353)
Cash	−1.072**	−1.069**	−1.078**	−1.083**	−1.071**	−1.064**
	(0.501)	(0.501)	(0.500)	(0.501)	(0.498)	(0.500)
Loss	4.000**	3.901**	3.902**	3.988**	3.911**	3.918**
	(1.698)	(1.693)	(1.692)	(1.700)	(1.691)	(1.693)
Big4	−0.468	−0.423	−0.446	−0.455	−0.439	−0.420
	(0.381)	(0.383)	(0.382)	(0.381)	(0.382)	(0.381)
Restate	−2.174***	−2.304***	−2.299***	−2.288***	−2.298***	−2.310***
	(0.605)	(0.615)	(0.617)	(0.616)	(0.615)	(0.617)
Year fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Cluster by firm	Yes	Yes	Yes	Yes	Yes	Yes
Observations	17,170	17,170	17,170	17,170	17,170	17,170
R-squared	0.5329	0.5325	0.5325	0.5326	0.5325	0.5326

Notes: This table presents the results of a comparison between our dictionary and five other dictionaries in predicting a firm's cyber risks in the following year. The detailed definition of variables can be found in Appendix A. The standard errors are clustered at the firm level and reported in parentheses. *, **, and *** denote significance at the 10 % , 5 % , and 1 % levels, respectively.

involved extracting and parsing text from these seed files. In Phase 3, we ranked the extracted words, and in Phase 4, we tested and assessed the dictionary's quality. This framework encourages researchers to think broadly about identifying relevant entities and documents while expanding sources beyond text to include audio and video recordings used by firms.

In the context of SEC investigations, our framework emphasizes the need to identify all entities, transactions, and relevant documents, including government laws. Firms' responses to SEC investigation letters can also guide subsequent content in CLs and may be included in the dictionary. Researchers interested in firms' strategies in response to SEC inquiries should consider firms' responses and remedial actions to prevent similar issues in the future. Our framework provides a customizable approach for developing dictionaries tailored to business and governance topics, contributing to research in text classification and information extraction. We validated our dictionary by applying it to textual analyses of SEC CLs to predict subsequent identification of ITCWs, IT audit fees, and cyber risks following SEC investigations. We compared its predictive performance against those of five benchmark dictionaries: Loughran and McDonald [17], the Harvard General Inquirer, the CFI financial dictionary, the Investopedia financial dictionary, and Henry [21]. Our dictionary demonstrated higher predictive accuracy than the others.

Our study makes three contributions to IS research. First, we developed a dictionary that is useful for IS research via a novel method we proposed in this study. SEC investigations offer a wealth of textual data relevant to a range of IT-related issues, yet this data source remains largely underexplored in IS research. For example, the SEC frequently investigates firms' IT controls and cybersecurity risk management, strategy, governance, and incidents. During its investigation process, the SEC may request information or documentation related to specific areas of concern to identify any violations or potential risks. The investigation may also involve interviews or examinations of IT systems or processes to gain a better understanding of the issues at hand [15,16]. Although interview transcripts and examination reports may not always be disclosed, when accessible, researchers can even transform audio and video data into textual data for analysis [58]. Despite the richness of these data, the literature has not effectively analyzed the content of these investigations because of the lack of a relevant high-quality dictionary [13]. Our dictionary can serve as an empirical tool to address this research gap, enabling future research to use the sentiment-devoid textual data for investigating IT-related issues considered pivotal by government and regulatory bodies.

Second, our study can facilitate research on ITCWs, IT audits, and cybersecurity risks. Our analysis builds on the work of C. Li et al. [3], who explored the consequences of ITCWs, including effects such as inaccurate management forecasts. Our dictionary effectively predicts ITCWs, addressing a critical gap in identifying weaknesses that might go unreported by firms or auditors. The ability to detect unreported ITCWs enables future studies to develop quantitative scores for assessing the likelihood of ITCWs in firms and to investigate factors contributing to the absence of reported weaknesses by auditors. These factors can include the role of auditors' IT-related expertise, which would offer a more comprehensive view of the IT audit process and cybersecurity risk management in firms. Moreover, we presented evidence suggesting that the content of SEC investigations influences firms' financial investments

Table 8
Robustness check: the weight of words.

Dependent variable	ITCW		IT_Audit_Fees		CyberRisk	
	(1)	(2)	(3)	(4)	(5)	(6)
MH_TFIDF	0.002***		0.001***		0.005**	
	(0.001)		(0.000)		(0.002)	
MH_Importance		0.003***		0.001***		0.006***
		(0.001)		(0.000)		(0.002)
LogAsset	0.662***	0.663***	0.009	0.010	0.092	0.090
	(0.035)	(0.036)	(0.009)	(0.009)	(0.129)	(0.129)
Leverage	0.090	0.090	0.331***	0.331***	3.175***	3.176***
	(0.064)	(0.064)	(0.041)	(0.041)	(0.330)	(0.330)
Current	0.017	0.018	0.103***	0.103***	2.042***	2.041***
	(0.126)	(0.126)	(0.025)	(0.025)	(0.455)	(0.455)
ROA	−0.151**	−0.150**	0.138***	0.138***	2.926***	2.927***
	(0.071)	(0.071)	(0.021)	(0.021)	(0.353)	(0.353)
Cash	0.046	0.049	−0.230***	−0.229***	−1.064**	−1.059**
	(0.138)	(0.138)	(0.036)	(0.036)	(0.501)	(0.501)
Loss	−0.623	−0.609	0.374**	0.377**	3.913**	3.927**
	(0.487)	(0.485)	(0.181)	(0.181)	(1.691)	(1.691)
Big4	0.021	0.025	0.053*	0.054*	−0.459	−0.457
	(0.201)	(0.201)	(0.029)	(0.030)	(0.380)	(0.381)
Restate	−0.572**	−0.572**	−0.170***	−0.170***	−2.240***	−2.220***
	(0.247)	(0.247)	(0.062)	(0.062)	(0.611)	(0.611)
Year fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Industry fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Cluster by firm	Yes	Yes	Yes	Yes	Yes	Yes
Observations	19,791	19,791	19,791	19,791	17,170	17,170
Pseudo-R-squared	0.3818	0.3823				
R-squared			0.3382	0.3383	0.5326	0.5327
Log pseudo-likelihood	−1,487.6930	−1,486.5152	–	–	–	–

Notes: This table presents the results of a robustness check for the constructs developed using our dictionary, weighted by TF-IDF and the feature importance of words. The detailed definition of variables can be found in Appendix A. The standard errors are clustered at the firm level and reported in parentheses. *, **, and *** denote significance at the 10 % , 5 % , and 1 % levels, respectively.

in IT audits. This highlighted the role of regulatory bodies in encouraging firms' commitment and efforts to enhance their technology governance, which has been overlooked in the literature [56]. Our dictionary's application in assessing the quality of firms' cybersecurity disclosures can significantly enhance understanding of the effects of cybersecurity disclosure quality. Although Gordon et al. [59] focused on how investors react to firms' cybersecurity disclosures, our study presented some evidence of the effects of the quality of these disclosures, not only their presence or absence. This approach allows researchers to delve deeper into the implications of firms' cybersecurity practices, particularly in cases in which chief information officers may misstate risks in cybersecurity management.

Third, our framework for dictionary development extends the ability of qualitative researchers to process a wide variety of qualitative content. The growing trend toward employing new data-collection methods has arisen from the now vast number of digital texts generated in everyday interactions, organizational practices, and online interventions [60]. This trend implies a need for new dictionaries to explore the word oceans in new problem domains. Given big data analytics, there is now an increased level of accessibility to various computational text-analysis approaches that can be employed by researchers who are not in the area of computer science, which creates exciting new opportunities for content-analysis research and allows scholars to access large, unstructured, and noisy textual (and other) data sources in unprecedented ways [61]. However, although some initial studies have outlined how computational textual analyses can be performed using various algorithms, these descriptions have often neglected to address, or only minimally addressed, the unique characteristics of the dictionary as the key input to such research.

By using our dictionary-construction approach, researchers can create dictionaries that are specific to one perspective of a problem (at the finer level) or the problem as a whole (at the broader level). Our approach also sheds light on how to advance techniques for natural language processing that focus on identifying sentiment polarity and shift the focus to specific information related to aspects or features (e.g., investigators' perspectives in our case) of an investigation dialogue, event, or entity mentioned in text. Given that identified keywords are specific to a context, using the constructed dictionary to model a problem is likely to yield better predictions. Moreover, highlighting the problem context in dictionary construction enables researchers to analyze data, identify patterns, and make predictions within their specific field of study more accurately and efficiently. This may also improve communication and collaboration between researchers and practitioners by providing a common language and understanding of technical terms and concepts. Our research represents an early step in shifting the focus of textual analysis from sentiment-oriented to action-oriented contextual analyses. Future research can use and extend our framework to develop high-quality dictionaries tailored to pivotal research questions in the field of IS, such as digital sustainability, health analytics, and information privacy. For example, researchers could create a dictionary to assess firms' compliance with the Digital Services Act or the ePrivacy Directive in Europe.

5.2. *Limitations and future research*

This study has three limitations that suggest directions for future research. First, our dictionary-construction framework was applied to a single dictionary focused on SEC investigations, using a bag-of-words approach that requires a fixed vocabulary size. This method can struggle with new terminology, often requiring re-analysis. Future research could address these limitations by exploring advanced techniques such as word embeddings (e.g., Word2Vec, GloVe) and deep learning models (e.g., transformers). In addition, expanding the framework to other IS domains, such as information privacy and digital healthcare, could create context-specific dictionaries based on domain-relevant seed files.

Second, while testing our dictionary across three contexts—including ITCWs, IT audit fees, and cyber risks—we relied on specific studies to operationalize constructs in each context. Although extensive testing

was conducted, variations in how variables were operationalized could still have influenced the results. Future studies could explore how operationalizations affect dictionary performance and assess the robustness of our findings across various approaches.

Third, the SEC, as a US regulatory body, produces documents with high standards of quality, making its CLs relatively easy to process. In contrast, documents from profit-driven firms, which prioritize revenue generation and market positioning, often vary in writing style, formality, and quality. Future research should investigate whether our dictionary-development method can be effectively applied to business and governance documents from small- and medium-sized firms in which less formal language and inconsistent quality may present greater challenges.

6. Conclusion

Our research underscores the importance of developing context-specific dictionaries for textual analysis, particularly sentiment-devoid data. By refining traditional dictionary-construction methods, we emphasized the significance of business communication in regulatory contexts. The SEC's documents offer stakeholders critical insights into a firm's financial and technology risks. Our tailored framework for governance documents outperformed widely used sentiment dictionaries, such as that of Loughran and McDonald [17]. Through empirical analysis, we demonstrated that our specialized dictionary more effectively captures the nuances of regulatory texts. This study highlights the necessity of context-specific dictionaries in analyzing business and governance materials and stresses the value of incorporating domain expertise in the dictionary-development process.

CRediT authorship contribution statement

Wentao Ma: Writing – review & editing, Writing – original draft, Formal analysis, Data curation, Conceptualization, Investigation, Methodology, Software, Validation, Visualization. **Shuk Ying Ho:** Supervision, Resources, Project administration, Conceptualization, Writing – review & editing.

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.

Acknowledgments

We thank two anonymous referees, an anonymous associate editor, and seminar participants at the Australian National University for their valuable feedback. We appreciate Tim Loughran and Bill McDonald for providing their dictionary, and we are grateful to Naaser Mohammed for his assistance.

Appendix A

Appendix A. Variable definitions

Variable	Definition	Source
Variables of interest		
MH	The count of words from our constructed dictionary that are found in the master letter, divided by the letter's total word count	EDGAR
LM	The count of negative sentiment words from the Loughran and McDonald [17] dictionary found in the master letter, divided by the letter's total word count	EDGAR
HarvardGI	The count of negative sentiment words from the Harvard General Inquirer dictionary that are found in the master letter, divided by the letter's total word count	EDGAR
CFI	The count of words from the CFI financial dictionary found in the master letter, divided by the letter's total word count	EDGAR
Investopedia	The count of words from the Investopedia financial dictionary that are found in the master letter, divided by the letter's total word count	EDGAR
Henry	The count of words from the Henry [21] dictionary found in the master letter, divided by the letter's total word count	EDGAR
MH_TFIDF	The sum of the term frequencies for words from our dictionary found in the master letter. Each is multiplied by the word's inverse document frequency. The term frequency of the word is measured as the count of the word that is found in the master letter, divided by the letter's total word count. The inverse document frequency of the word is measured as one plus the natural logarithm of the total number of master letters in our sample, divided by the total number of master letters that contain the word	EDGAR
MH_Importance	The sum of counts for words from our dictionary that are found in the master letter. Each is weighted by the word's feature importance and multiplied by the inverse document frequency of the word	EDGAR
Dependent variables		
RV_RES	An indicator that equals 1 if the firm makes any financial restatements before the completion of the SEC investigation, and 0 otherwise	Audit Analytics
ITCW	An indicator that equals 1 if the firm reports any ITCW in the year following the SEC investigation, and 0 otherwise	Audit Analytics
IT_Audit_Fees	The value of IT audit fees paid to auditors in the year following the SEC investigation	Audit Analytics
CyberRisk	The predicted value of the model of Asthana et al. [57] in the year following the SEC investigation. The model is as follows: $CyberBreach_{i,t+1} = \beta_0 + \beta_1 LogAsset_{i,t} + \beta_2 Sales_{i,t} + \beta_3 ROA_{i,t} + \beta_4 ICW_{i,t} + \beta_5 ZScore_{i,t} + \beta_6 BtoM_{i,t} + \beta_7 Leverage_{i,t}$ + IndustryFE+ YearFE+ $\varepsilon_{i,t}$. *CyberBreach* is an indicator that equals 1 if the firm discloses a cyber breach in the year following the SEC investigation, and 0 otherwise. *Sales* is measured as the natural logarithm of firm sales. *ZScore* is measured as 0.3 × net income/total assets+ sales/total assets+ 1.4 × retained earnings/total assets+ 1.2 × working capital/total assets+ 0.6 × marker value/total liabilities. *BtoM* is measured as the book-to-market equity ratio of the firm	Audit Analytics and Compustat
Control variables		
ICW	The number of internal control weaknesses unrelated to IT that the firm reports	Audit Analytics
LogAsset	The natural logarithm of total assets	Compustat
Leverage	Total liabilities, scaled by total assets	Compustat
Current	Current assets	Compustat
ROA	Net income, scaled by total assets	Compustat
Cash	Cash, scaled by total assets	Compustat
Loss	An indicator that equals 1 if the firm reports negative net income in the year following the SEC investigation, and 0 otherwise	Compustat

(*continued on next page*)

(*continued*)

Variable	Definition	Source
Big4	An indicator that equals 1 if the firm is audited by a Big 4 auditor in the year following the SEC investigation, and 0 otherwise	Compustat
Restate	An indicator that equals 1 if the firm makes any financial restatements in the year following the SEC investigation, and 0 otherwise	Audit Analytics

References

[1] T. Jetzek, M. Avital, N. Bjorn-Andersen, The sustainable value of open government data, J. Assoc. Inf. Syst. 20 (6) (2019) 702–734, https://doi.org/10.17705/1jais.00549.

[2] M.S. Pang, G.H. Lee, The impact of IT decision-making authority on IT project performance in the US federal government, MIS Q. 46 (3) (2022) 1759–1776, https://doi.org/10.25300/MISQ/2022/16898.

[3] C. Li, G.F. Peters, V.J. Richardson, M.W. Watson, The consequences of information technology control weaknesses on management information systems: the case of Sarbanes–Oxley internal control reports, MIS Q. 36 (1) (2012) 179–203, https://doi.org/10.2307/41410413.

[4] J.P. Ryans, Textual classification of SEC comment letters, Rev. Account. Stud. 26 (1) (2021) 37–80, https://doi.org/10.1007/s11142-020-09565-6.

[5] X. Fang, J. Zhan, Sentiment analysis using product review data, J. Big. Data 2 (2015) 5, https://doi.org/10.1186/s40537-015-0015-2. Article.

[6] S.Y. Deng, A.P. Sinha, H.M. Zhao, Adapting sentiment lexicons to domain-specific social media texts, Decis. Support. Syst. 94 (2017) 65–76, https://doi.org/10.1016/j.dss.2016.11.001.

[7] A. Madsen, S. Reddy, S. Chandar, Post-hoc interpretability for neural NLP: a survey, ACM. Comput. Surv. 55 (8) (2023) 155, https://doi.org/10.1145/3546577. Article.

[8] M. Boukes, B. van de Velde, T. Araujo, R. Vliegenthart, What's the tone? Easy doesn't do it: analyzing performance and agreement between off-the-shelf sentiment analysis tools, Commun. Methods Meas. 14 (2) (2019) 83–104, https://doi.org/10.1080/19312458.2019.1671966.

[9] W. Van Atteveldt, M.A. Van der Velden, M. Boukes, The validity of sentiment analysis: comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms, Commun. Methods Meas. 15 (2) (2021) 121–140, https://doi.org/10.1080/19312458.2020.1869198.

[10] W.Y. Li, S.Y. Phang, K.W. Choi, S.Y. Ho, The strategic role of CIOs in IT controls: IT control weaknesses and CIO turnover, Inform. Manage. 58 (6) (2021) 103429, https://doi.org/10.1016/j.im.2021.103429. Article.

[11] O. Turel, Q.H. He, Y.T. Wen, Examining the neural basis of information security policy violations: a noninvasive brain stimulation approach, MIS Q. 45 (4) (2021) 1715–1744, https://doi.org/10.25300/MISQ/2021/15717.

[12] J. Haislip, J.H. Lim, R. Pinsker, The impact of executives' IT expertise on reported data security breaches, Inform. Syst. Res. 32 (2) (2021) 318–334, https://doi.org/10.1287/isre.2020.0986.

[13] L.M. Cunningham, J.J. Leidner, The SEC filing review process: a survey and future research opportunities, Contemp. Account. Res. 39 (3) (2022) 1653–1688, https://doi.org/10.1111/1911-3846.12742.

[14] P.M. Dechow, A. Lawrence, J.P. Ryans, SEC comment letters and insider sales, Account. Rev. 91 (2) (2016) 401–439, https://doi.org/10.2308/accr-51232.

[15] T.W. Wang, J.C. Yen, K. Yoon, Responses to SEC comment letters on cybersecurity disclosures: an exploratory study, Int. J. Account. Inform. Syst. 46 (2022) 100567, https://doi.org/10.1016/j.accinf.2022.100567. Article.

[16] P. Rosati, F. Gogolin, T. Lynn, Audit firm assessments of cyber-security risk: evidence from audit fees and SEC comment letters, Int. J. Account. 54 (3) (2019) 1950013, https://doi.org/10.1142/S1094406019500136. Article.

[17] T. Loughran, B. McDonald, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, J. Finance 66 (1) (2011) 35–65, https://doi.org/10.1111/j.1540-6261.2010.01625.x.

[18] P.J. Stone, R.F. Bales, J.Z. Namenwirth, D.M. Ogilvie, The general inquirer: a computer system for content analysis and retrieval based on the sentence as a unit of information, Behav. Sci. 7 (4) (1962) 484, https://doi.org/10.1002/bs.3830070412. Article.

[19] T. Vipond, Financial analysis ratios glossary, 2024. https://corporatefinanceinstitute.com/resources/accounting/financial-analysis-ratios-glossary/.

[20] Investopedia, Dictionary, 2024. https://www.investopedia.com/financial-term-dictionary-4769738.

[21] E. Henry, Are investors influenced by how earnings press releases are written? J. Bus. Commun. 45 (4) (2008) 363–407, https://doi.org/10.1177/0021943608319388.

[22] A. Bodnaruk, T. Loughran, B. McDonald, Using 10-K text to gauge financial constraints, J. Financ. Quant. Anal. 50 (4) (2015) 623–646, https://doi.org/10.1017/S0022109015000411.

[23] R. Frankel, J. Jennings, J. Lee, Disclosure sentiment: machine learning vs. dictionary methods, Manage. Sci. 68 (7) (2022) 5514–5532, https://doi.org/10.1287/mnsc.2021.4156.

[24] S.Y. Ho, K.W. Choi, F. Yang, Harnessing aspect-based sentiment analysis: how are tweets associated with forecast accuracy? J. Assoc. Inf. Syst. 20 (8) (2019) 1174–1209, https://doi.org/10.17705/1jais.00564.

[25] S.Y. Deng, Z.J. Huang, A.P. Sinha, H.M. Zhao, The interaction between microblog sentiment and stock returns: an empirical examination, MIS Q. 42 (3) (2018) 895–918, https://doi.org/10.25300/MISQ/2018/14268.

[26] I. Dissanayake, S. Nerur, R. Singh, Y. Lee, Medical crowdsourcing: harnessing the "wisdom of the crowd" to solve medical mysteries, J. Assoc. Inf. Syst. 20 (11) (2019) 1589–1610, https://doi.org/10.17705/1jais.00579.

[27] X.M. Luo, B. Gu, J. Zhang, C.W. Phang, Expert blogs and consumer perceptions of competing brands, MIS Q. 41 (2) (2017) 371–396, https://doi.org/10.25300/MISQ/2017/41.2.03.

[28] Y.C. Chang, C.H. Ku, D.D.L. Nguyen, Predicting aspect-based sentiment using deep learning and information visualization: The impact of COVID-19 on the airline industry, Inform. Manage. 59 (2) (2022) 103587, https://doi.org/10.1016/j.im.2021.103587. Article.

[29] Z. Shaukat, A.A. Zulfiqar, C. Xiao, M. Azeem, T. Mahmood, Sentiment analysis on IMDB using lexicon and neural networks, SN. Appl. Sci. 2 (2) (2020) 148, https://doi.org/10.1007/s42452-019-1926-x. Article.

[30] B. Burscher, D. Odijk, R. Vliegenthart, M. de Rijke, C.H. de Vreese, Teaching the computer to code frames in news: comparing two supervised machine learning approaches to frame analysis, Commun. Methods Meas. 8 (3) (2014) 190–206, https://doi.org/10.1080/19312458.2014.937527.

[31] W. Guo, M. Diab, Semantic topic models: combining word distributional statistics and dictionary definitions, in: R. Barzilay, M. Johnson (Eds.), Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 552–561. https://aclanthology.org/D11-1051.

[32] X. Li, H. Xie, L. Chen, J. Wang, X. Deng, News impact on stock price return via sentiment analysis, Knowl. Based. Syst. 69 (2014) 14–23, https://doi.org/10.1016/j.knosys.2014.04.022.

[33] Dictionsoftware. (2024). *What is diction?* https://dictionsoftware.com/diction-overview/.

[34] A.K. Davis, I. Tama-Sweet, Managers' use of language across alternative disclosure outlets: earnings press releases versus MD&A, Contemp. Account. Res. 29 (3) (2012) 804–837, https://doi.org/10.1111/j.1911-3846.2011.01125.x.

[35] T. Loughran, B. McDonald, Textual analysis in accounting and finance: a survey, J. Account. Res. 54 (4) (2016) 1187–1230, https://doi.org/10.1111/1475-679X.12123.

[36] D.H. Solomon, E. Soltes, D. Sosyura, Winners in the spotlight: media coverage of fund holdings as a driver of flows, J. Financ. Econ. 113 (1) (2014) 53–72, https://doi.org/10.1016/j.jfineco.2014.02.009.

[37] H.L. Chen, P. De, Y. Hu, B.H. Hwang, Wisdom of crowds: the value of stock opinions transmitted through social media, Rev. Financ. Stud. 27 (5) (2014) 1367–1403, https://doi.org/10.1093/rfs/hhu001.

[38] J. Engelberg, M. Henriksson, A. Manela, J. Williams, The partisanship of financial regulators, Rev. Financ. Stud. 36 (11) (2023) 4373–4416, https://doi.org/10.1093/rfs/hhad029.

[39] P.T.Y. Lee, E. Feiyu, M. Chau, Defining online to offline (O2O): a systematic approach to defining an emerging business model, Internet Res. 32 (5) (2022) 1453–1495, https://doi.org/10.1108/INTR-10-2020-0563.

[40] C. Kearney, S. Liu, Textual sentiment in finance: a survey of methods and models, Int. Rev. Financ. Anal. 33 (2014) 171–185, https://doi.org/10.1016/j.irfa.2014.02.006.

[41] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.

[42] X.W. Huang, D. Kroening, W.J. Ruan, J. Sharp, Y.C. Sun, E. Thamo, M. Wu, X.P. Yi, A survey of safety and trustworthiness of deep neural networks: verification, testing, adversarial attack and defence, and interpretability? Comput. Sci. Rev. 37 (2020) 35, https://doi.org/10.1016/j.cosrev.2020.100270. Article.

[43] T. Ito, K. Tsubouchi, H. Sakaji, T. Yamashita, K. Izumi, Contextual sentiment neural network for document sentiment analysis, Data Sci. Eng. 5 (2) (2020) 180–192, https://doi.org/10.1007/s41019-020-00122-4.

[44] N.T.M. Demoulin, K. Coussement, Acceptance of text-mining systems: the signaling role of information quality, Inform. Manage. 57 (1) (2020) 103120, https://doi.org/10.1016/j.im.2018.10.006. Article.

[45] J.C. Zimmer, R.E. Arsal, M. Al-Marzouq, V. Grover, Investigating online information disclosure: effects of information relevance, trust and risk, Inform. Manage. 47 (2) (2010) 115–123, https://doi.org/10.1016/j.im.2009.12.003.

[46] R. Lukyanenko, J. Parsons, Y.F. Wiersma, M. Maddah, Expecting the unexpected: effects of data collection design choices on the quality of crowdsourced user-generated content, MIS Q. 43 (2) (2019) 623–647, https://doi.org/10.25300/MISQ/2019/14439.

[47] A. Al-Hroob, A.T. Imam, R. Al-Heisa, The use of artificial neural networks for extracting actions and actors from requirements document, Inf. Softw. Technol. 101 (2018) 1–15, https://doi.org/10.1016/j.infsof.2018.04.010.

[48] P. Gupta, A. Sharma, R. Jindal, Scalable machine-learning algorithms for big data analytics: a comprehensive review, Wiley Interdiscipl. Rev. Data Mining Knowl. Discov. 6 (6) (2016) 194–214, https://doi.org/10.1002/widm.1194.
[49] K. Kheiri, H. Karimi, SentimentGPT: Exploiting GPT for advanced sentiment analysis and its departure from current machine learning, *arXiv* (Preprint) (2023). https://arxiv.org/abs/2307.10234.
[50] Y. Zhang, P. Tino, A. Leonardis, K. Tang, A survey on neural network interpretability, IEEe Trans. Emerg. Top. Comput. Intell. 5 (5) (2021) 726–742, https://doi.org/10.1109/TETCI.2021.3100641.
[51] K. Yang, R.Y.K. Lau, A. Abbasi, Getting personal: a deep learning artifact for text-based measurement of personality, Inform. Syst. Res. 34 (1) (2022) 194–222, https://doi.org/10.1287/isre.2022.1111.
[52] M. Ghiassi, D. Zimbra, S. Lee, Targeted Twitter sentiment analysis for brands using supervised feature engineering and the dynamic architecture for artificial neural networks, J. Manage. Inform. Syst. 33 (4) (2016) 1034–1058, https://doi.org/10.1080/07421222.2016.1267526.
[53] M. Kraus, S. Feuerriegel, Decision support from financial disclosures with deep neural networks and transfer learning, Decis. Support. Syst. 104 (2017) 38–48, https://doi.org/10.1016/j.dss.2017.10.001.
[54] N. Mahmoudi, P. Docherty, P. Moscato, Deep neural networks understand investors better, Decis. Support. Syst. 112 (2018) 23–34, https://doi.org/10.1016/j.dss.2018.06.002.
[55] D. Stoel, D. Havelka, J.W. Merhout, An analysis of attributes that impact information technology audit quality: A study of IT and financial audit practitioners, Int. J. Account. Inform. Syst. 13 (1) (2012) 60–79, https://doi.org/10.1016/j.accinf.2011.11.001.
[56] T. Mazza, S. Azzali, Information technology controls quality and audit fees: evidence from Italy, J. Account. Audit. Finance 33 (1) (2018) 123–146, https://doi.org/10.1177/0148558X15625582.
[57] S.C. Asthana, R. Kalelkar, K. Raman, Does client cyber-breach have reputational consequences for the local audit office? Account. Horiz. 35 (4) (2021) 1–22, https://doi.org/10.2308/HORIZONS-2020-018.
[58] E. Aguirre, A.L. Roggeveen, D. Grewal, M. Wetzels, The personalization–privacy paradox: implications for new media, J. Consum. Market. 33 (2) (2016) 98–110, https://doi.org/10.1108/JCM-06-2015-1458.
[59] L.A. Gordon, M.P. Loeb, T. Sohail, Market value of voluntary disclosures concerning information security, MIS Q. 34 (3) (2010) 567–594, https://doi.org/10.2307/25750692.
[60] T.M. Paulus, J.N. Lester, Doing qualitative research in a digital world, Sage Publications, 2021.
[61] F. Muller-Hansen, M.W. Callaghan, J.C. Minx, Text as big data: develop codes of practice for rigorous computational text analysis in energy social science, Energy Res. Soc. Sci. 70 (2020) 101691, https://doi.org/10.1016/j.erss.2020.101691. Article.
[62] I. Sheikh, I. Illina, D. Fohr, G. Linares, Learning word importance with the neural bag-of-words model. Proceedings of the 1st Workshop on Representation Learning for NLP, Association for Computational Linguistics, 2016, pp. 222–229, https://doi.org/10.18653/v1/W16-1626.
[63] E. Ngai, P.T.Y. Lee, A review of the literature on applications of text mining in policy making, in: C. Hsu, J.Y.L. Thong, S.X. Xu, M.D. Marco, M. Limayem (Eds.), Proceedings of the 20th Pacific Asia Conference on Information Systems, Association for Information Systems, 2016, pp. 1–12. https://aisel.aisnet.org/pacis2016/343.
[64] P.T.Y. Lee, R.W.C. Lui, M. Chau, B.H.Y. Tsin, Exploring the effects of different achievement goals on contributor participation in crowdsourcing, Inform. Technol. People 36 (3) (2023) 1179–1199, https://doi.org/10.1108/ITP-08-2020-0583.

Wentao Ma is an Assistant Professor at the International Business School at the Xi'an Jiaotong-Liverpool University. His research interests cover the impact of big data on businesses. His publications have appeared in *Information & Management.* and *Finance Research Letters*.

Shuk Ying Ho is a Professor of Information Systems at the Australian National University. Her research interests cover personalization, electronic government, open-source software development, social media, data analytics, and accounting information systems. Her publications have appeared in *MIS Quarterly, Information Systems Research, Journal of the Association for Information Systems, European Journal of Operational Research, Decision Support Systems, Journal of Business Ethics*, and many others. She is serving on the editorial boards of *MIS Quarterly, Journal of Management Information Systems,* and *Journal of the Association for Information Systems*.