Journal of Accounting Research

DOI: 10.1111/1475-679X.12294
Journal of Accounting Research
Vol. 58 No. 1 March 2020
Printed in U.S.A.

CHICAGO BOOTH

What Are You Saying? Using *topic* to Detect Financial Misreporting

NERISSA C. BROWN,* RICHARD M. CROWLEY,† AND W. BROOKE ELLIOTT*

Received 12 September 2016; accepted 25 October 2019

ABSTRACT

We use a machine learning technique to assess whether the thematic content of financial statement disclosures (labeled *topic*) is incrementally informative in predicting intentional misreporting. Using a Bayesian topic modeling algorithm, we determine and empirically quantify the topic content of a large collection of 10-K narratives spanning 1994 to 2012. We find that the algorithm produces a valid set of semantically meaningful topics that predict financial misreporting, based on samples of Securities and Exchange

*University of Illinois at Urbana-Champaign; †Singapore Management University.

Accepted by Phil Berger. We thank an anonymous reviewer, Andrew Bauer, Matt Cobabe, Amanda Convery, Robert Davidson, Paul Demeré, Lucile Faurel, Shawn Gordon, Jing He, Shiva Rajgopal, Kristina Rennekamp, Kecia Williams Smith, Gang Wang, and workshop participants at Baruch College—City University of New York, Carnegie Mellon University, Columbia University, Hong Kong University of Science and Technology, Nagoya University, University of Illinois, U.S. Securities and Exchange Commission (Division of Economic and Risk Analysis), Virginia Tech, the 2015 AAA FARS Mid-year Meeting, the 2015 AAA Annual Meeting, the 2015 Conference on Convergence of Financial and Managerial Accounting Research, the 2016 Conference on Investor Protection, Corporate Governance, and Fraud Prevention, and the 2016 Conference on Financial Economics and Accounting for helpful comments. We also thank Xiao Yu for insightful comments on methodology and coding, Brian Gale for helpful assistance with Amazon Mechanical Turk, and Stephanie Grant, Chunlei Liu, Jill Santore, and Jingpeng Zhu for excellent research assistance. We thank Derryck Coleman and Olga Usvyatsky (formerly) of Audit Analytics for assistance with the restatement data and text search scripts used in this study. Brown gratefully acknowledges financial support from the PricewaterhouseCoopers LLP Faculty Fellowship. Elliott gratefully acknowledges financial support from the Ernst & Young Distinguished Professorship. An online appendix to this paper can be downloaded at http://research.chicagobooth.edu/arc/journal-of-accounting-research/online-supplements.

Commission (SEC) enforcement actions (Accounting and Auditing Enforcement Releases [AAERs]) and irregularities identified from financial restatements and 10-K filing amendments. Our out-of-sample tests indicate that *topic* significantly improves the detection of financial misreporting by as much as 59% when added to models based on commonly used financial and textual style variables. Furthermore, models that incorporate *topic* significantly outperform traditional models when detecting serious revenue recognition and core expense errors. Taken together, our results suggest that the topics discussed in annual report filings and the attention devoted to each topic are useful signals in detecting financial misreporting.

JEL codes: C80; K22; K42; M40; M41; M48

Keywords: topic modeling; disclosure; latent Dirichlet allocation; financial misreporting

1. Introduction

This study investigates whether a novel text-based measure of the thematic content of financial statement disclosures (labeled as *topic*) is useful for detecting financial misreporting.[1] Detection models have long focused on quantitative financial statement and stock market variables as predictive factors (Beneish [1997], Brazel, Jones, and Zimbelman [2009], Dechow et al. [2011], Bao et al. [2020]). One drawback of this approach is that financial misreporting can go undetected for multiple periods, because misreporting firms often manipulate performance metrics and accounting transactions to blend in better with their peers or the firm's own past performance (Lewis [2013]). To address this weakness, recent studies analyze the textual and linguistic features of management disclosures, finding that summary measures of these features serve as useful warnings of misreporting (see, e.g., Hobson, Mayew, and Venkatachalam [2012], Larcker and Zakolyukina [2012], Purda and Skillicorn [2015]).

Despite the usefulness of communication style in revealing misreporting, the literature debates whether textual and linguistic features adequately capture managers' deliberate attempts to obfuscate or manipulate financial information (Bloomfield [2008], Bushee, Gow, and Taylor [2018]). Further, as Loughran and McDonald [2016] highlight, commonly used textual measures do not reflect the context or meaning of management disclosures, thereby limiting the inferences that can be drawn. We tackle these issues by introducing a machine learning tool that simultaneously detects and quantifies the thematic content (*topic*) of annual report

[1] We use the terms *misreporting* and *misrepresentation* interchangeably to refer to deliberate violations of financial accounting standards and noncompliance with regulatory financial reporting rules. We refrain from using the term *fraud* because, in a legal sense, violations of or noncompliance with financial reporting standards and rules are considered fraudulent only if market participants rely on the misreported or misrepresented information to their detriment.

narratives. This approach departs from prior text-based research by focusing on *what* is being disclosed by management rather than *how*. Using this unique measure, we evaluate the disclosure topics associated with misreporting and how these topics evolve. More importantly, we investigate the incremental predictive power of *topic* in detecting misreporting out of sample, relative to a collection of financial and textual style measures.

Our focus on the thematic content of financial statement filings draws on the management disclosure and communications literatures. These bodies of research suggest that the flexible nature of disclosure content allows for a broader set of dimensions along which annual report narratives can be used to identify financial misreporting, compared to quantitative financial metrics and summary measures of textual features (Hoberg and Lewis [2017]). These literatures also argue that textual features, such as tone and word usage, are difficult to classify as deceptive, because disclosure narratives can be influenced by individuals' expectations and motivations, even when the intent is to communicate objectively and truthfully (Douglas and Sutton [2003]). In that sense, the content of the disclosure and the attention devoted to each topic may better predict misreporting than how the narrative is fashioned. We therefore examine whether the topic content of financial statement disclosures is incrementally informative in assessing the likelihood of misreporting, beyond textual style features. We also analyze the ability of *topic* to detect misreporting, relative to quantitative financial variables, given that these measures are typically backward-looking and less efficient in predicting misreporting, compared to language-based measures (Cecchini et al. [2010a], Goel and Gangolly [2012], Larcker and Zakolyukina [2012], Purda and Skillicorn [2015]).

We generate our *topic* measure by employing a Bayesian topic modeling algorithm developed by Blei, Ng, and Jordan [2003], termed Latent Dirichlet Allocation (LDA). Similar to factor or cluster analysis, LDA is an unsupervised and unstructured probabilistic model that "learns" or discovers the latent thematic structure of words within a corpus of documents.[2] The algorithm (and other variants) is widely used in practice by Internet search engines to guide keyword selection and improve correlations between search terms and web content (Fishkin [2014]). A unique advantage of LDA is that it does not require predetermined word dictionaries or topic categories and instead relies on the fact that words frequently appearing together tend to be semantically related. This process reduces researcher bias, as foreknowledge of document content does not affect the topic classifications.[3] Furthermore, the algorithm can classify the content of large

[2] LDA is a "bag of words" algorithm that uses the distribution of words across documents to classify and quantify themes without the need for predefined or researcher-determined word lists or topic categories.

[3] Although LDA is unsupervised and does not rely on human input to identify topics, human judgment is necessary to interpret and label the topics inferred from the algorithm. This is because the LDA output for a given topic consists only of word clusters and word

collections of textual narratives—a task that would be otherwise impractical with large samples of financial statement filings.

We derive our *topic* measure using a comprehensive sample of 131,528 10-K filings issued by U.S. firms from 1994 to 2012 (based on the electronic filing date). The full text of each 10-K filing is retrieved from the Securities and Exchange Commission's (SEC) EDGAR system and parsed following the procedures in Li [2008]. We run the LDA algorithm on the parsed filings using rolling five-year windows over our sample period. This time-series approach allows the topic categories to evolve, as we expect time-varying factors to influence management communications as well as the ability of thematic content to detect financial misreporting. The topics discovered in each five-year window are then used to compute the proportion of each topic discussed in 10-K filings issued in the subsequent year.

We use three data sources to identify instances of intentional misreporting. This multiple-pronged sampling approach follows from Karpoff et al. [2017, KKLM], who demonstrate the importance of evaluating empirical inferences drawn from different data sources of financial misreporting. Our first data source is a comprehensive database of SEC Accounting and Auditing Enforcement Releases (AAERs), developed by Dechow et al. [2011] and compiled by the University of California at Berkeley's Center for Financial Reporting and Management (CFRM). The data set provides details on firms subject to SEC enforcement actions for alleged accounting or auditing violations and the pertinent reporting periods. We use this information to identify the 10-K filings affected by the violations.

We rely on financial restatements and amended annual filings (10-K/As) as two additional sources of intentional misreporting. We broadly refer to these forms of intentional misreporting as *irregularities*. We gather restatements from Audit Analytics (AA) and capture irregularities by screening the data for restatements categorized as "fraudulent" or "nonclerical" or those associated with a regulatory investigation, following the classification method in Karpoff et al. [2017, KKLM]. We conduct an automated text search of 10-K/As for material misrepresentations or omissions that are seemingly intentional. Using the criteria in Hennes, Leone, and Miller [2008, HLM], we classify the original 10-K filing as misreported if our search tool identifies the following in the filing amendment: (1) variants of the words "fraud" or "irregularity" in describing the filing revision and (2) references to revisions stemming from investigations by the SEC, U.S. Department of Justice (DOJ), or independent parties. Although this process relies on a single disclosure source, the irregularities gathered from 10-K/As are valuable, as they capture a broad set of financial reporting problems.

There is significant overlap between our three data sources, though each sample has advantages and disadvantages. Misreporting events drawn from

---

probabilities (also referred to as word weightings). Thus, some amount of discretion is unavoidable when applying LDA.

AAERs provide researchers with high confidence of intentional misreporting, because the SEC targets firms when there is strong evidence of intent to mislead (Dechow et al. [2011], KKLM). Indeed, a benchmarking analysis by KKLM finds that AAERs fare the best in capturing price-moving events related to financial misconduct. The AAER data also outperform restatement samples in identifying infractions that meet the statutory definition of securities fraud. However, one drawback is that many instances of misreporting are not pursued by the SEC, due to resource constraints (Files [2012]). Another shortcoming is that cases pursued may reflect selection biases arising from the SEC's evaluation and investigation processes. Our irregularity samples mitigate these limitations, because the events are drawn from broader sources. These samples could, however, introduce other selection or identification issues as the procedures for gathering these data depend on how firms disclose and discuss misreporting events in filings.[4]

We first evaluate the semantic validity of the LDA output and the ability of the topic categories to detect misreporting in samples on an annual basis. Using both human and machine-based procedures, we find that LDA produces a coherent set of semantically meaningful topics that capture the economic content of annual report filings. These topics capture content referring to firm performance, complex transactions, financing arrangements, and risk factors. Interestingly, the discussion of particular topics evolves, indicating that the content of management communications is quite fluid. Our in-sample tests indicate that several topics are consistently associated with misreporting across our three samples. For instance, we find that misreporting firms devote more attention to discussing increases in financial performance and business activities, such as mergers and acquisitions (M&As) and share capital transactions, that arguably create incentives to misreport (Hoberg and Lewis [2017]). Misreporting firms, on the other hand, allocate less attention to discussing issues related to risk factors, cost commitments, and credit agreements.

Our next analyses assess the usefulness of *topic* in detecting misreporting out of sample, compared to a comprehensive set of quantitative financial and textual style variables. The financial variables stem from an expanded version of the Dechow et al. [2011] *F-score* model, whereas the textual variables (denoted *Style*) cover various measures of textual complexity, language voice, and disclosure tone. Our detection process is real time, as we apply a rolling five-year prediction approach, similar to the time-series method used to generate the LDA topics. We find that *topic* provides significant incremental predictive power over our collection of *F-score* and *Style* variables. Models that incorporate *topic* outperform standalone or joint models of *F-score* and *Style*, depending on the type of misreporting event.

[4] A limitation that is common to all three samples is that each data source may provide an incomplete history or sequence of events that constitutes a case of misreporting (KKLM).

When we evaluate the interplay between all three sets of predictors, we find that a joint model of *topic* and *F-score* best predicts misreporting in the AAER and AA samples, whereas a joint model of *topic* and *Style* best predicts misreporting in the 10-K/A sample. This differential result indicates that disclosure topics and financial metrics complement each other in detecting misreporting that prompts an SEC action or a restatement, whereas topic- and style-based information is more useful in identifying broad financial reporting problems.

Additional tests of the classification accuracy of our models indicate that *topic* improves the classification of high-risk cases of misreporting in the AAER (AA) samples by as much as 59% (50%) when added to models based solely on financial and textual style variables. The economic value of *topic* is more modest in our 10-K/A sample, with a 6% increase in classification accuracy for high-risk cases. We also find that *topic* significantly improves the detection of serious revenue-recognition and core-expense errors. Sensitivity checks indicate that our results are unaffected by fluctuations in misreporting over time and are not driven by firms with repeated misreporting events. Our inferences are also robust to an expanded set of financial and textual style measures and to topics derived solely from the Management Discussion and Analysis (MD&A) section of the annual report.

Our study makes several important contributions. First, we extend the literature by documenting that the topics discussed in annual report filings are useful for identifying intentional misreporting, either on a standalone basis or in combination with standard prediction variables. Second, we expand burgeoning research that examines the thematic content of financial disclosure narratives (e.g., Dyer, Lang, and Stice-Lawrence [2017], Huang et al. [2018]). We exploit a robust machine learning tool that quantifies *what* is being disclosed in annual financial reports (as opposed to *how*). This content analysis is a significant step forward, as it delves deeper into the semantic meaning of management disclosures and how disclosure themes point to misreporting. Further, our real-time prediction method considers the time-varying and fluid nature of management communications, which contrasts with prior work based on word dictionaries, which are fairly static and easily identifiable by firms.

Lastly, our study has important practical implications for regulators and corporate monitors, who have begun to implement text-based initiatives aimed at detecting financial misrepresentation. For instance, the SEC has developed computer-powered risk assessments that leverage text-based tools, such as word dictionaries and topic modeling, to detect anomalies in firm disclosures (Eaglesham [2013], Bauguess [2018]). Auditors are similarly employing textual analytics to identify reporting risks (Murphy and Tysiac [2015]). Our study suggests that extracting disclosure themes from corporate filings helps capture financial misdeeds.[5]

[5] Regulators and monitors should note that topic analysis is less susceptible to "gaming," relative to other text-based measures, as the words in any topic category and the associated

2. *Background and Research Questions*

2.1 PREDICTING FINANCIAL MISREPORTING

Over the past two decades, researchers have sought to identify a parsimonious set of predictors of financial misreporting. Research documents that measures of extreme or abnormal financial performance are useful predictors of misreporting. For instance, studies find that misreporting firms exhibit high abnormal accruals, disproportionate increases in receivables and inventories, and poor abnormal market performance (Feroz, Park, and Pastena [1991], Beneish [1997, 1999]). In addition, studies suggest that misreporting is associated with stock and debt market pressures (Dechow, Sloan, and Sweeney [1996]) and weaknesses in firms' internal governance or monitoring (Beasley [1996], Beasley et al. [2000], Farber [2005]).

Drawing from this literature, Dechow et al. [2011] conduct a comprehensive analysis of the quantitative financial and stock characteristics of misreporting firms identified in AAERs. They find that AAER firms have low-quality accruals, deteriorating financial and nonfinancial performance, and high stock valuations, relative to fundamentals. AAER firms are also more likely to engage in aggressive off-balance-sheet and financing activities during the misstatement period. Using these characteristics, Dechow et al. [2011] develop a composite prediction measure (*F-score*) that outperforms traditional accrual measures in detecting misreporting. However, despite this added performance, studies argue that the predictive ability of quantitative characteristics is quite modest (Larcker and Zakolyukina [2012]), with many of the measures behaving opposite to conventional wisdom (Purda and Skillicorn [2015]). To address this weakness, recent research explores the predictive value of various language-based measures. The premise is that the linguistic features of management disclosures reveal communication patterns that foretell financial misreporting. This literature relies on two general approaches to uncover and analyze communication patterns in written disclosures (see Li [2010b] and Loughran and McDonald [2016] for reviews of these methodologies.)

The first approach relies on predefined word categorizations (or dictionaries) to investigate the link between intentional misreporting and language tone as well as deception cues. For instance, Li [2008] finds that the relative frequency of self-reference words, causation words, positive emotion words, and future tense verbs within the MD&A section of 10-K filings is associated with managerial obfuscation. Loughran and McDonald [2011] find that negative and uncertain language in 10-K filings is linked to securities lawsuits of alleged accounting improprieties, whereas Rogers, Buskirk, and Zechman [2011] report that firms sued for financial

---

weightings are part of an interdependent system. Thus, gaming topic analysis would require a complex and evolving unraveling system.

misreporting use more optimistic language in their earnings announcements. Finally, Larcker and Zakolyukina [2012] analyze conference call transcripts and find that deceptive language, such as emotion words and anxiety words, better predicts misreporting, compared to abnormal accrual measures.

The second approach employs machine learning algorithms to discriminate between "bags of words" or textual style markers that predict intentional misreporting. These markers include textual features, such as verbal complexity, readability, and disclosure tone, as well as grammar and word choices. Among these studies, Cecchini et al. [2010a], Goel et al. [2010], and Purda and Skillicorn [2015] are most relevant to our research. All three studies use a supervised learning algorithm, termed the Support Vector Machine (SVM), to classify misreporting events. The SVM tool improves upon prior work, as it learns by example and does not require predefined language markers. For instance, Purda and Skillicorn [2015] apply an SVM to detect misreporting based on word usage in both annual and quarterly filings. They find that an SVM-generated word dictionary outperforms prediction models built using predefined dictionaries as well as models based on quantitative financial measures.

Despite advances in the literature regarding the use of language-based cues and textual markers to detect financial misreporting, there is little research that incorporates the deeper semantic meaning of management communications when assessing the likelihood of misreporting. Simply put, there is scant evidence on whether what managers choose to discuss predicts misreporting. Our study addresses this gap in the literature by using a state-of-the-art topic analysis tool to capture the thematic content of annual report narratives.

2.2 LDA TOPIC MODELING

We employ a topic modeling approach developed by Blei, Ng, and Jordan [2003], termed LDA, to capture the thematic content of annual financial reports. The LDA technique is widely used in the linguistic and information retrieval literatures to identify the thematic structure of text corpora and other collections of discrete disclosure data. (See Blei [2012] for a review of topic modeling and its applications.) We use this approach to construct a firm-specific measure (*topic*) of the topics discussed in annual financial statements in a given reporting year. This unique measure (defined as the normalized percentage of the annual report attributed to each topic identified by the algorithm) captures the extent to which a particular topic is discussed within a given annual report filing.

Topic modeling is relatively new to accounting and finance (see Loughran and McDonald [2016] and Eickhoff and Neuss [2017] for literature reviews), and our measurement approach is consistent with recent studies that use LDA to investigate various financial reporting and capital market phenomena. For instance, LDA has been used to examine (1) the link between stock market movements and topics frequently searched by

internet users (Curme et al. [2014]), (2) the thematic content of analyst reports and conference call narratives (Huang et al. [2018]), (3) the types of risks discussed in 10-K filings and how these discussions influence investors' risk perceptions (Bao and Datta [2014]), and (4) the topics that contribute to increases in 10-K length over time (Dyer, Lang, and Stice-Lawrence [2017]).

More closely related to our study, Hoberg and Lewis [2017] use LDA to examine whether misreporting firms provide abnormal MD&A disclosures and the underlying incentives for this behavior. Hoberg and Lewis [2017] find that, relative to industry peers, AAER firms disclose abnormal content that is common among misreporting firms. Their topic analysis indicates that AAER firms are more likely to grandstand manipulated revenue and R&D expense performance during misreporting years, while providing fewer quantitative details about their financial results. This evidence is consistent with incentives for misreporting firms to tout strong performance and growth options, while concealing broad performance details. AAER firms also under disclose topics related to liquidity challenges, perhaps to mitigate negative cost-of-capital effects. Lastly, Hoberg and Lewis [2017] find that managers disassociate themselves from the firm during misreporting years by disclosing abnormally low levels of content referencing their participation in the firm's vision and strategy.

Our study differs from Hoberg and Lewis [2017] in several respects. First, we move beyond disclosure incentives to demonstrate the incremental predictive power of thematic content in detecting misreporting within the broader population of financial statement filers. As a result, our study seeks to provide new insights into the benefits of statistical topic analysis in assessing the likelihood of misreporting. Second, Hoberg and Lewis [2017] fit their LDA model using annual reports filed in only the first year of their sample period (1997). This approach does not account for changes in disclosure topics over time and likely induces staleness in the topics used in their analyses. As highlighted elsewhere (Cecchini et al. [2010a], Li [2010b], Dyer, Lang, and Stice-Lawrence [2017]), accounting for temporal variations in managerial communication is an important and much needed extension of current text-based methodologies. We therefore extend Hoberg and Lewis [2017] by employing a rolling-window estimation procedure that accounts for the time-varying nature of the topics discussed by management. We use the same rolling-window setup to predict misreporting in real time using the topics identified in the window immediately preceding the prediction period.

Lastly, the analysis of Hoberg and Lewis [2017] is confined to the MD&A section, whereas our study considers the thematic content of the entire 10-K filing. Although the MD&A section provides a useful setting for examining disclosure content, it does not capture topics discussed elsewhere in the annual report (Li [2010b]). Moreover, a major drawback of focusing on only one section of the annual report is that companies can strategically

shift or (de-)emphasize content across multiple sections (Loughran and McDonald [2016]).[6]

2.3 RESEARCH QUESTIONS

Our research questions explore the *incremental* value of thematic content in identifying intentional misreporting, relative to the predictive value of quantitative financial and textual style characteristics. Although research has yet to examine the detection power of disclosure topics, several arguments in the literature suggest that the thematic content of financial statement filings is likely to capture an aspect of managerial deception that is distinct from what can be gleaned from firms' financial characteristics or the textual style of the filings.

Regulatory oversight is more difficult for financial statement narratives, especially at the topic level, thus leaving more room for managers to use disclosure content to obfuscate or mislead. Although researchers have identified a set of deceptive words and language cues from disclosure narratives (Newman et al. [2003], Larcker and Zakolyukina [2012]), it is more difficult for regulators to identify and monitor a set of deceptive topics naïvely, as these topics may be benign or informative about true financial performance in other settings or at other times. Furthermore, deceptive managers have discretion to vary not only the topics discussed in financial statement filings but also the amount of attention devoted to each. Thus, as argued by Hoberg and Lewis [2017], the flexible nature of disclosure content allows for a broader set of dimensions along which financial statement narratives can be used to identify intentional misreporting.

Research finds that a sizable proportion of the textual narratives in earnings releases and annual report filings contain forward-looking statements that cover a wide range of topics (Li [2010a], Bozanic, Roulstone, and Van Buskirk [2018]). This is in marked contrast to quantitative financial information, which is primarily backward looking. Evidence also suggests that management's discussion of forward-looking information responds differently to firm conditions, such as poor earnings performance and financial distress, compared to discussions of backward-looking information (Bonsall, Bozanic, and Merkley [2014]). As such, it appears reasonable to expect that financial statement topics will respond differently to managerial deception and obfuscation, relative to backward-looking quantitative metrics.

The detection value of disclosure topics beyond textual style is yet another important query, as linguistics research suggests that it is difficult to discern deception or obfuscation from the textual features of disclosure

[6] Consistent with this argument, Amel-Zadeh and Faasse [2016] find that the tone of footnote disclosures is more negative than that of the MD&A section, especially when firm performance is poor. This result likely reflects managers' attempt to downplay negative information by putting a more positive spin on MD&A content. A similar shifting strategy might be at play in our setting as extended analyses show that MD&A topics have lower detection power, compared to topics identified from the entire 10-K filing.

narratives. In fact, this research shows that communication features, such as tone and abstract language, can be flavored by individuals' expectations and motivations, even when the intent is to communicate objectively and truthfully (Douglas and Sutton [2003]). Accounting researchers also debate whether textual features, such as length, readability, and word usage, reflect managerial deception or obfuscation or simply the inherent complexity of discussing unusual performance or business events (Bloomfield [2008], Loughran and McDonald [2016]). Indeed, evidence suggests that textual complexity is an ambiguous indicator of managerial obfuscation that often reflects informative technical disclosures or complex accounting and regulatory standards (Guay, Samuels, and Taylor [2016], Dyer, Lang, and Stice-Lawrence [2017], Bushee, Gow, and Taylor [2018]).[7]

Overall, the above arguments suggest that annual report topics may improve the detection of financial misreporting, beyond what can be achieved using quantitative financial metrics and aggregate measures of textual style features. However, any incremental detection power is an empirical question, given the variety of ways that deception and obfuscation can manifest in written communications. In addition, financial statement filings are joint outputs of management and legal counsel, leading to content that is vague and boilerplate in some areas (Brown and Tucker [2011], Hoberg and Lewis [2017]). This conservative aspect of financial reporting further implies need for an agnostic set of research questions, as follows.

Research Question 1 (RQ1): Do disclosure topics improve the detection of intentional financial misreporting, relative to quantitative financial measures?

Research Question 2 (RQ2): Do disclosure topics improve the detection of intentional financial misreporting, relative to aggregate textual style features?

3. Data and Empirical Measures

3.1 DATA AND SAMPLE SELECTION

We base our topic analysis on the textual narratives contained in annual 10-K filings. We focus on 10-K filings, as opposed to other firm disclosures, because they (1) provide comprehensive coverage of the firm and its activities throughout the fiscal year, (2) avoid selection biases given their mandatory status, and (3) maximize the number of firm-year observations in our prediction tests. We download the full text of all 10-Ks available through the SEC EDGAR FTP site from January 1, 1994 (the first year of available

[7] Similar arguments can be made for the topics communicated in firm disclosures because thematic content can reflect business events or regulatory standards as opposed to managerial obfuscation. Nonetheless, in contrast to summary textual measures, the detailed nature of topic analysis allows for greater insight into how disclosure content and the attention management devoted to each topic vary with misreporting.

data), until December 31, 2012. The download yields 131,528 10-Ks filed by U.S. firms over the period. We use the full set of filings to generate the disclosure topics, as this improves the algorithm's convergence.

We parse the 10-K filings using the approach of Li [2008], but expand the methodology to remove all items in the filings other than raw text. We describe our parsing methodology in appendix A.1 of the online appendix. After generating our topic and textual style measures, we merge the sample of 10-K filings with Compustat and CRSP, from which we gather financial-statement and stock-market data, respectively. We exclude financial firms (Standard Industrial Classifications [SIC] codes 6000-6799) and those firm-years with missing Compustat and CRSP data. We gather data on intentional misreporting from three separate data sources as discussed below and then merge the 10-K filing sample with each data source.

3.1.1. Identifying Intentional Financial Misreporting. We use three data sources to identify instances of intentional financial misreporting.[8] Our first data source relies on SEC AAERs to identify misreporting firms and the affected filings. The AAER data were originally compiled by Dechow et al. [2011], and were most recently updated by the University of California at Berkeley's CFRM. The updated data set includes all AAERs issued by the SEC on or before September 30, 2016, and covers accounting and auditing violations that affect filings issued prior to and including the 2012 calendar year. We create an indicator variable (*misreport*) that equals 1 for each annual reporting period affected by an accounting or auditing violation as identified in the enforcement release and zero otherwise.[9] We then use the *misreport* variable to classify the corresponding 10-K filing as misreported. From the AAER sample, we identify 505 misreported 10-K filings for 192 unique firms issued from 1994 to 2010. The AAER data do not provide sufficient violation events that map to our sample in 2011 and 2012.

We draw our next set of misreporting events from the AA Non-Reliance Restatements database, which covers restatements announced or disclosed from January 1, 2001, onward. These restatements map to misreported 10-K filings issued during the latter part of our sample period (2000 through 2012). Unlike the AAER data set, the AA data provide a mix of

[8] Alternative sources include (1) the Government Accountability Office (GAO) restatement database, (2) restatement data collected by HLM from Form 8-K filings, and (3) data on SEC and DOJ enforcement actions for Rule 13(b) violations as used in KKLM and provided by Call et al. [2018]. We cannot incorporate these data sets into our analyses, as they do not identify the fiscal period in which the misreporting occurred, or the filing that was affected. Nonetheless, our samples overlap with these alternative sources, because they rely on similar underlying data. Research also shows that our data sources have fewer limitations regarding sample coverage and identification of the misreporting period (see KKLM for comparisons of the AAER, AA, GAO, and 13(b) data sets).

[9] Consistent with Dechow et al. [2011], we exclude enforcement actions for misconduct that is unrelated to accounting or auditing (e.g., bribery or disclosure-related violations). We also exclude enforcement actions that do not identify a specific reporting period in which the violation occurred.

restatements arising from both unintentional errors and intentional violations. Given our focus on intentional misreporting, we use the categorization process outlined by KKLM to cull the data of unintentional errors. This process is based on the HLM classification scheme and uses three criteria to identify intentional misreporting, referred to as irregularities: (1) the use of the words "fraud" or "irregularity" (or their variants) in reference to the restatement, (2) whether the restatement is associated with an SEC or DOJ investigation, and (3) the presence of an independent investigation in relation to the restatement.

Based on the above criteria, KKLM rely on four variables in the AA database to distinguish irregularities from errors. We follow the same approach and classify a restatement as an irregularity if the restatement disclosure references financial fraud, irregularities, or misrepresentations (AA variable *res_fraud* = 1), or indicates the involvement of the SEC, P, or other regulator in the restatement process (*res_sec_investigation* = 1).[10] We also treat a restatement as an irregularity if the misstatement is neither a simple misapplication of a GAAP or FASB accounting rule (*res_accounting* = 0) nor an accounting or clerical error (*res_clerical_errors* = 0). For each irregularity, we identify the 10-K filing containing the misstatement and code *misreport* as 1. This process yields 527 misreported filings issued by 245 unique firms from 2000 through 2012.[11]

Our final misreporting sample is based on a customized automated search of amended 10-K filings for material misrepresentations or disclosure omissions that are seemingly intentional. Filing amendments provide clear advantages in identifying misreporting, compared to conventional restatement samples. Amended 10-Ks are routinely filed over our entire sample period (1994–2012) and thus provide a longer history of irregularities, relative to the AA restatement sample.[12] They also capture financial reporting and internal control problems that affect not only the financial statements but also the footnote disclosures, the auditor's report, and management's report on internal controls over financial reporting.

We download all 10-K/A filings from the SEC EDGAR FTP site and gather firm-identifying information for matching purposes from the header (or alternately from the body of the text when the header is missing or

---

[10] The AA data set offers a broader scope of SEC involvement than outlined in the second HLM criterion. Specifically, the AA taxonomy includes restatements involving an SEC comment letter and those arising from a PCAOB investigation or an investigation by another regulator.

[11] Seventy-three percent of our AA irregularities are nonreliance ("Big R") restatements that require the reissuance of the financial statements. Thus, our classification scheme maps strongly with accounting corrections that are deemed material by management.

[12] The AA data also rely on 10-K/As as a source of identifying financial restatements. We cross check the two samples over the 2000–2012 overlapping period and find that 35% of the misreported filings identified from 10-K/As map to a restated filing in the full AA database. This rate of overlap is reasonable, given that AA only archives amendments that correct line items reported on the face of the financial statements.

incomplete). We then parse the 10-K/As, using the same method applied to the original 10-K filings. Next, we apply an automated text search procedure (see appendix A), based on the HLM classification criteria and a list of keywords recommended by AA (2012), to identify irregularities from disclosure narratives. We code *misreport* as 1 for each 10-K that our search string identifies as containing an irregularity in the corresponding 10-K/A. We set *misreport* to 0 if there is no amended 10-K for the respective filing year or if the 10-K/A filing does not reveal an irregularity. This process yields 697 misreported filings across 553 unique firms from 1994 through 2012.

A manual inspection of 50 randomly selected observations indicates that our customized search tool performs well in capturing material filing revisions that fit the HLM criteria.[13] We hand check another random sample of 20 10-K/As that were not flagged as an irregularity by our search string. We find that only 1 of the 20 observations reflects an irregularity, indicating that false negatives are not a major concern. We check a final random sample of 20 10-K/As flagged as irregularities and find that 76% of the cases coincide with auditor resignations or dismissals, internal control failures, late filings, SEC comment letters, and CFO resignations in the current or subsequent reporting year.[14] Thus, amended 10-K filings seem to provide strong signals of financial reporting and compliance problems.

Table 1 reports the frequency of financial misreporting by calendar year for all three data sources, along with the overall percentage of firm-years with detected misreporting events. Panel A presents the frequencies for the AAER sample, while panels B and C present the frequencies for the AA and 10-K/A samples, respectively. For each panel, we report the percentage of misreporting events across all sample years and for the years included in our out-of-sample prediction tests. The prediction years appear below the horizontal line in each panel. Consistent with prior research (Larcker and Zakolyukina [2012], Perols et al. [2017], Bao et al. [2020]), the rate of misreporting is very low, ranging from 1.34% to 1.65% across all years in the three samples and from 1.18% to 1.89% across the prediction years. The rate of misreporting in the AAER sample declines substantially after 2005. This evidence is consistent with Bao et al. [2020], who document that the AAER sample is more stable prior to 2005, due to the long time lag in SEC investigations and possible changes in the SEC's enforcement priorities after the 2008 financial crisis. We observe a similar post-2005 decline

[13] We randomly draw 50 of the amended 10-Ks that are flagged as disclosing an irregularity by our search tool. Of the 50 observations, 44 (88%) capture material revisions that fit the HLM criteria. These cases include (1) restatements resulting from internal investigations, reaudits, or SEC comment letters; (2) revisions to footnote disclosures to correct errors related to revenue recognition and core expense accounts; and (3) material omissions from the MD&A and risk factor disclosures. The remaining six observations pertain to cases where the 10-K/A filing insufficiently explains the revision. We conduct a further screen to remove amendments that are viewed as more technical. (See appendix A.6.2 of the online appendix.)

[14] We use the Accounting Quality + Risk Matrix in AA to identify these events.

TABLE 1
Distribution of Financial Misreporting

	Panel A: AAERs			Panel B: AA Irregularities			Panel C: 10-K/A Irregularities		
Year	Observations	Frequency	Percentage	Observations	Frequency	Percentage	Observations	Frequency	Percentage
1994	786	0	0.00				786	2	0.25
1995	1,043	6	0.58				1,043	2	0.19
1996	1,634	17	1.04				1,634	17	1.04
1997	2,250	23	1.02				2,250	16	0.71
1998	2,308	40	1.73				2,308	12	0.52
1999	2,195	47	2.14				2,195	13	0.59
2000	2,041	51	2.50	2,041	36	1.76	2,041	21	1.03
2001	2,021	44	2.18	2,021	39	1.93	2,021	18	0.89
2002	2,391	51	2.13	2,391	60	2.51	2,391	27	1.13
2003	2,936	60	2.04	2,936	81	2.76	2,936	53	1.81
2004	2,843	52	1.83	2,843	77	2.71	2,843	70	2.46
2005	2,678	39	1.46	2,678	75	2.80	2,678	65	2.43
2006	2,608	19	0.73	2,608	39	1.50	2,608	78	2.99
2007	2,549	18	0.71	2,549	19	0.75	2,549	53	2.08
2008	2,535	13	0.51	2,535	14	0.55	2,535	48	1.89
2009	2,564	15	0.59	2,564	28	1.09	2,564	65	2.54
2010	2,424	10	0.41	2,424	24	0.99	2,424	48	1.98
2011				2,330	19	0.82	2,330	44	1.89
2012				2,178	16	0.73	2,178	45	2.07

(*Continued*)

TABLE 1—*Continued*

	Panel A: AAERs			Panel B: AA Irregularities			Panel C: 10-K/A Irregularities		
Year	Observations	Frequency	Percentage	Observations	Frequency	Percentage	Observations	Frequency	Percentage
All firm-years	37,806	505	1.34	32,098	527	1.64	42,314	697	1.65
No. of firms	6,423	192		5,082	245		6,588	553	
Prediction firm-years	29,785	419	1.41	19,866	234	1.18	34,293	648	1.89
No. of firms	5,259	162		3,916	148		5,427	513	

This table reports the frequency and percentage of financial misreporting events by year for each of our three data sources. Panel A presents the yearly frequencies for the AAER sample, whereas panels B and C present the frequencies for the Audit Analytics (AA) and 10-K/A irregularity samples, respectively. Each sample excludes firm observations with missing data in Compustat and CRSP to compute the set of quantitative financial statement and stock return variables used in our prediction models. The last four rows of each panel report overall frequencies and percentages for all sample years and for the out-of-sample prediction years. The prediction years for the AAER sample (panel A) extends from 1999 through 2010 and excludes the 2011 and 2012 years due to the lack of misreporting events that map to our sample. The prediction years for the AA sample (panel B) extend from 2005 through 2012 because the AA restatement database is incomplete prior to 2000. The prediction years for the 10-K/A sample (panel C) extend from 1999 through 2012. The first five-year estimation window for the AAER and 10-K/A samples runs from 1994 through to 1998, whereas the first estimation window for the AA sample runs from 2000 through 2004.

in the AA sample. This decline has been attributed to improvements in internal control assessments and financial reporting quality after the passage of the 2002 Sarbanes Oxley Act (SOX), combined with stronger oversight of firm disclosures in the SEC's filing review process (Choudhary, Merkley, and Schipper [2019]).

3.2 EMPIRICAL MEASURES

3.2.1. Financial and Textual Style Measures. We draw our quantitative financial statement and stock market variables from the Dechow et al. [2011] *F-score* model (see model 3 in table 9 of their paper). These variables capture accrual quality, firm performance, off-balance-sheet activities, and market pressures. We augment the *F-score* model with variables capturing firm size, audit quality, and involvement in complex business transactions, namely, M&As and restructurings. These additional variables capture characteristics that are correlated with reporting risks and the quality of firms' external and internal monitoring mechanisms (Farber [2005], Doyle, Ge, and McVay [2007], Ashbaugh-Skaife et al. [2008]). Panel A of appendix B defines all of our financial variables.

We compute a comprehensive set of textual features (denoted *Style*) as described in panel B of appendix B. These measures include common proxies for readability, textual complexity, and disclosure tone (Li [2008], Loughran and McDonald [2011]). We also capture deeper linguistic markers, such as language voice, lexical variety, and disclosure emphasis (Goel et al. [2010], Goel and Gangolly [2012]), Rennekamp [2012], Purda and Skillicorn [2015]). Finally, we construct four additional measures derived from our parsing process: (1) the log of the number of bullets, (2) the length of the SEC mandated header, (3) the number of excess newlines (vertical white space), and (4) the character length of HTML tags. These measures are exploratory and attempt to control for unobservable factors.

3.2.2. LDA Topic Measure. Our measure of thematic content (*topic*) is based on the unstructured and unsupervised LDA topic modeling methodology developed by Blei, Ng, and Jordan [2003]. We choose this approach due to its intuitive characteristics and strong performance. LDA is a Bayesian probabilistic model and offers significant improvements over older data-driven and principle-component–based tools, such as Latent Semantic Analysis (LSA). Further, the topic modeling accuracy of LDA is quite strong, when compared to human topic classification. For example, Anaya [2011] finds that humans classify topics with 94% accuracy, whereas LDA achieves 84% accuracy. Likewise, using human judges, Chang et al. [2009] find that LDA produces semantically meaningful and coherent topics that correspond well to human concepts.

In the context of business narratives, Eickhoff and Neuss [2017] report that LDA and other topic models have been successfully applied to textual documents in several disciplines, including accounting, finance,

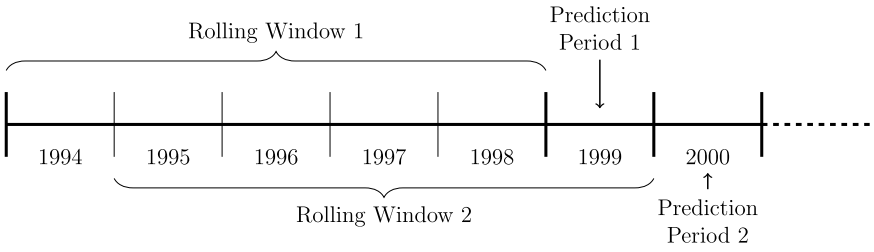
information systems, and management.[15] Many of these studies use qualitative or quantitative methods (or both) to evaluate the effectiveness of LDA in identifying business-related topics. Qualitatively, Bellstam et al. [2019] find that LDA generates topics from analyst reports that conform well to concepts used to describe innovation activities. Quantitative analyses further confirm that these topics are strongly correlated with innovation measures. Huang et al. [2018] use a human coder to classify the topics discussed in the conference call transcripts and analyst reports for three firms in the same industry. They find that the LDA topic assignments match the topics identified by the human coder about two-thirds of the time. Using qualitative techniques, such as topic labeling and visual analysis of economic trends, Huang et al. [2018] also find that LDA reliably captures the economic content of analyst reports and conference calls. Lastly, Dyer, Lang, and Stice-Lawrence [2017] employ the human validation technique of Chang et al. [2009] and find that LDA topics derived from 10-K narratives are coherent and meaningful when tested against human intuition. Collectively, this body of evidence indicates that LDA is effective at classifying the thematic content of business-related narratives.

The LDA model is based on a few simple assumptions. It assumes a collection of K topics in a given document and that the vocabulary of each topic is distributed following a Dirichlet distribution, $\beta_K \sim \text{Dirichlet}(\eta)$. The model further assumes that the topic proportions in each document d are drawn from a Dirichlet distribution $\theta_d \sim \text{Dirichlet}(\alpha)$. Given these assumptions, a specific number of topics to identify, and a few learning parameters, the LDA model categorizes the words in a given set of documents into well-defined topics. Because the model uses Bayesian analysis, a word is allowed to be associated with multiple topics. This is a distinguishing feature of LDA, as words can have multiple meanings, especially in different contexts. In short, LDA is a probabilistic process that condenses the vocabulary in a collection of documents into a dictionary of topics and a set of topic weights.

We implement LDA using a dynamic time-series process, because annual report content is likely to vary over time, due to macroeconomic or industry trends, changes in disclosure requirements, or managerial turnover (Dyer, Lang, and Stice-Lawrence [2017]). This approach allows us to assess the changing nature of disclosure content and its ability to predict misreporting over time. Our time-series procedure identifies the topics discussed in 14 rolling five-year windows over our sample period (1994 through 2012). That is, we run the algorithm for the periods 1994–1998, 1995–1999, 1996–2000, and so on. The topics discovered in each five-year window are then used to determine the disclosure content of annual reports issued in the

[15] One of the first studies to apply topic modeling to business documents is by Boukus and Rosenberg [2006], who use LSA to analyze the content of the Federal Open Market Committee's (FOMC) meeting minutes.

Panel A: Rolling Window Test Procedure



Panel B: Treatment of Misreporting Events

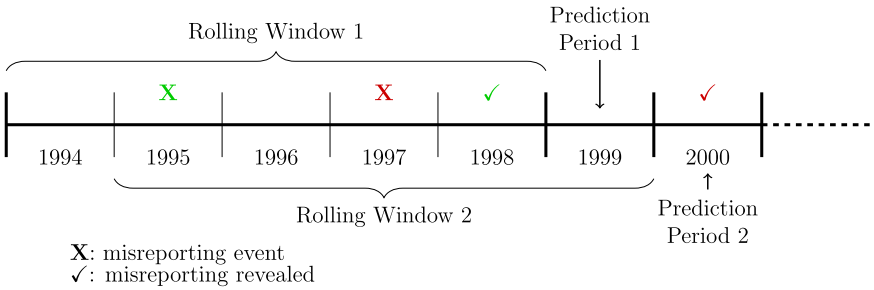


FIG. 1.—Time-series research design. This figure illustrates the setup of our rolling time-series estimation and prediction procedures. Panel A depicts the general structure of our rolling five-year estimation periods and the prediction years immediately following each window. Panel B depicts our treatment of misreporting events within each window for the samples of irregularities drawn from Audit Analytics (AA) and amended 10-K filings (10-K/As). Misreporting events are included in the in-sample estimation window *only if* the event is publicly revealed by the end of the five-year window. Misreporting events that are revealed outside of the five-year period are excluded from the in-sample estimation period. As illustrated in panel B, the first misreporting event (green X) is included in both estimation periods (1994–1998 and 1995–1999), whereas the second misreporting event (red X) is excluded. The process described above does not apply to the AAER sample because the enforcement release dates are not available in the CFRM database. Thus, all misreporting events in the AAER sample are presumed to be publicly known by the end of each estimation window.

year immediately following each window. This results in a test period of 1999–2012 for our prediction analyses. (This period is reduced to 1999–2010 for the AAER sample and to 2005–2012 for the AA sample.) Panel A of figure 1 illustrates the setup of our rolling window analysis. Note that, although new topics may arise in the year after each window, the topics discussed in the prior five years provide the most practical estimates of current-year disclosure content while avoiding potential look-ahead biases.

We follow Hoffman, Bach, and Blei [2010] and implement LDA using an online batch variant of the algorithm. This approach is computationally efficient as the algorithm is applied to small batches at a time (100 filings).

We draw the filings in each batch in random order to mitigate overweighting of early years in the online LDA tool. We then set the algorithm to identify 31 topics in each five-year window. We select 31 topics because simulated results indicate that this number of topics is optimal in detecting irregularities drawn from amended 10-K filings.[16] We run the LDA algorithm on the filings, generating 31 topics in each window and the weighting for each word associated with the topic. We use these word weights to compute the weighting of each topic in filings issued in the year after the window (e.g., the word weights for topics identified in the 1994 to 1998 window are applied to filings issued in 1999). The topic weights in a given filing are computed by multiplying the vector of word weights for each topic by a vector of word counts from the filing. We normalize the topic weights by scaling by the sum of the weights of all topics identified in the filing. This procedure generates the proportion of the annual report narrative (ranging from 0 to 1) that is devoted to each topic. We further orthogonalize the topic proportions to 2-digit SIC to adjust for unobserved industry effects.[17] We denote the industry-normalized topic proportions as *topic*.

4. Empirical Results

4.1 EVALUATION OF LDA TOPIC MEASURE

We assess the validity of the LDA topics before tackling our research questions. Because LDA is unsupervised, it is necessary to evaluate the algorithm's effectiveness in capturing human intuitions. We follow prior studies and use both human and machine-based evaluation methods to assess the semantic meaning and interpretability of the topics inferred from the 10-K narratives (Chang et al. [2009], Quinn et al. [2010], Bao and Datta [2014], Dyer, Lang, and Stice-Lawrence [2017], Huang et al. [2018], Bellstam et al. [2019]).[18]

---

[16] We run the simulation on our 10-K/A sample, given the lack of misreporting events for multiple years in the AAER and AA samples. The details of our simulations are in appendix A.2 of the online appendix.

[17] Our results are unchanged when we also orthogonalize the topic measure to audit firm type (Big *N*, midsize, or small firm) to control for unobserved audit firm effects.

[18] An alternative approach is the use of human coders to classify topics discussed in a subset of our 10-K narratives. The topics identified by the human coders would then serve as a benchmark for evaluating the LDA topic assignments. We refrain from using this approach, given the time-consuming and rigid nature of manual coding (Quinn et al. [2010]). To make human classification tasks feasible, studies typically present coders with a fixed list of potential topic labels prior to commencing the task. The coders would then read units of text and attempt to assign one of the topics from the fixed list to each unit. This directed form of human coding is used in both Anaya [2011] and Huang et al. [2018] as a benchmarking method. (See section 3.2.2 for earlier discussion of these studies.) Although ex ante knowledge of potential topics is a reasonable design choice, one drawback is that this approach is likely to bias or limit human judgments, thereby reducing the usefulness of human classification as a validation method. We therefore focus on evaluating whether our LDA topics have human-identifiable semantic

4.1.1. Interpretation and Labeling of LDA Topic Output. Our first method evaluates the semantic meaning of the LDA output by labeling the topics and assessing the extent to which they provide meaningful economic content. This form of evaluation is qualitative and follows from several studies, including those by Quinn et al. [2010], Bao and Datta [2014], Bellstam et al. [2019], Hoberg and Lewis [2017], and Huang et al. [2018]. However, one limitation of this approach is that it naturally involves human discretion, given the manual process of topic labeling. This limitation, nonetheless, is not a concern for our empirical analyses, as our prediction tests are based on the *quantitative* topic proportions within each filing *and not* the assigned labels.

As discussed above, we derive our topic measure using a time-series approach, with 31 topics identified in each of the 14 rolling five-year windows over our sample period. For ease of interpretation and labeling, we aggregate the topics discovered in each window up to the full sample. We refer to the aggregate topics as "combined topics." We allow multiple topics within a given window to be associated with the same combined topic. We also allow the number of combined topics to exceed 31, as some topics do not appear in every window. We derive the combined topics by matching topics across years based on the Pearson correlation of the word weights within the topics. All topics with a Pearson correlation above a specific threshold are grouped together. We test correlation thresholds from 1% to 90% in 1% intervals to determine the most coherent grouping. The most coherent grouping is achieved at the 11% threshold, resulting in 64 combined topics across our sample period.

To determine the underlying content of each combined topic, we generate a list of the highest weighted phrases and sentences associated with each topic. (See Hoberg and Lewis [2017] and Dyer, Lang, and Stice-Lawrence [2017] for a similar approach.) We construct the list by first extracting the top 1,000 sentences per topic based on the weighted words associated with each combined topic. Next, we sort the sentences based on length and extract the middle tercile (334 sentences) as representative sentences of typical length. The 20 most frequent bigrams (two-word phrases excluding stop words, numbers, and symbols) are then extracted from the 334 midlength sentences. These sentences are also sorted based on the cosine similarity between a given sentence and the remaining 333 sentences. We manually review the top 20 bigrams and top 100 midlength sentences based on cosine similarity, and assign descriptive labels to each of the combined topics.

Appendix A.3 of the online appendix lists the inferred topic labels for the 64 combined topics.[19] We note that the LDA algorithm performs well

---

coherence. This approach is superior, in that it does not require devising a list of benchmark topics or providing our LDA topic labels to human judges beforehand.

[19] The topic labels for the first 50 topics are also available in figures 2–4. Appendix A.3 also presents 10 representative bigrams and two representative midlength sentences for

in identifying narrative content that relates distinctively to changes in firms' financial performance and their financing activities. For instance, combined topics 1 and 2 both refer to income performance, compared to prior periods. Examples of top midlength sentences from both topics are "The company's gross profit margin decreased to 59% in Fiscal 1996, compared to 65% in Fiscal 1995" and "Operating profit was $122.8 million in 2011, compared with $113.9 million in 2010, an increase of 7.8%." Other performance- and financing-related topics include segment performance (combined topics 12 and 40), debt issuances and credit arrangements (19, 33, and 45), and equity transactions (39). LDA also identifies topics related to complex business transactions such as hedging activities (8 and 29) and M&As and spin-offs (11, 22, 30, and 38).

Several topics refer to specific financial statement items and their underlying measurement or recognition assumptions. These include accounts receivable and doubtful accounts (10), long-term assets (16), revenue recognition (17), advertising expenses (25), and postretirement cost assumptions (3). Consistent with Bao and Datta [2014], we find that LDA discovers content related to business risks and contingencies, such as foreign currency risks (43), country risks (13 and 26), and litigation (31).[20] Lastly, the algorithm identifies several industry-specific topics (50–64), though we do not discuss them for brevity's sake.

4.1.2. Word Intrusion Tasks. Our next evaluation method uses "word intrusion" tasks to assess the semantic coherence of the *unaggregated* topics derived by the algorithm across each of the rolling windows. Chang et al. [2009] argue that the overall interpretability of LDA-derived topics can be evaluated by the extent to which human subjects agree with the makeup of the topics. Using this logic, they develop a word intrusion task in which human subjects attempt to identify an unrelated or "intruder" word inserted into a list of words that LDA selects as belonging to the same topic. If the set of words from the LDA model is coherent, then the human subjects should easily identify the intruder word at a rate that is significantly higher than random chance. Thus, a higher identification rate indicates higher interpretability of the LDA output.

We conduct our word intrusion tasks using both human-subject as well as machine-based procedures. Given the many topics across our rolling

---

each topic. The reported bigrams exclude redundant phrases and those with similar inferences. Note also that the inferred labels for a few topics are overlapping, due to only minor differences in the content inferred from the bigrams and midlength sentences. This overlap does not affect our empirical tests, as we estimate our prediction models using the individual LDA topics and not the combined topics.

[20] We follow Bao and Datta [2014] and assess whether our LDA topics are correlated with investors' risk perceptions as proxied by future stock return volatility. Consistent with their evidence, we find that 10-K discussions of macro-level risk factors, namely, foreign currency and country risks, environmental risks, and commodity risks, are positively associated with future return volatility. We also find that discussions of derivative and hedging activities, securitizations, and business collaborations influence risk perceptions.

windows, we are limited in using human subjects to test the coherence of all the topics inferred by LDA. Our machine-based procedures mitigate this limitation by allowing us to test the coherence of the entire set of topics. We briefly discuss the setup of the task procedures and our results and provide more detailed descriptions in appendix A.4 of the online appendix.

We take all of the unaggregated topics discovered within each rolling window and randomly select 3 of the 10 most probable words associated with each topic based on the word weights from the LDA model. An intruder word is then selected at random from a pool of the top 10 words appearing in another random topic. We next apply a word embedding algorithm to test all possible word combinations across all the topics (i.e., three topic words plus a random intruder word). Our human-subjects procedure follows from Chang et al. [2009] and is conducted using a short experimental task with 180 online workers on Amazon Mechanical Turk (MTurk). We make the MTurk task practicable by asking participants to identify the intruder word from a small subset (20) of the word combinations evaluated by the machine algorithm. The results from both procedures indicate identification rates that are statistically higher than random chance (25% or one out of four words) at the 1% level. The machine-based procedure correctly identifies the intruder word with accuracy rates ranging from 50% to 53%, while the human-subjects task produces an average accuracy rate of 40%.

Taken together, our qualitative and quantitative evaluation methods suggest that the LDA algorithm provides a valid set of semantically meaningful topics that are reasonably coherent and interpretable by human judges.

4.2 PREDICTIVE VALUE OF LDA TOPIC MEASURE

4.2.1. Empirical Methodology. We investigate our research questions by estimating in-sample prediction models over rolling five-year windows. (See panel A of figure 1.) We then conduct out-of-sample tests using the regression estimates from each five-year window to predict the likelihood of intentional misreporting in the year after the end of each window.

For filings coded as misreported in the AA and 10-K/A irregularity samples, we ensure that the misreporting event is revealed by the end of the in-sample window. That is, the in-sample prediction model for a given five-year window incorporates a misreporting event *only if* the event is publicly known by the end of the window. Misreporting events that are revealed outside of the five-year window are excluded from the in-sample model (*misreport* is recoded as zero for the respective filing in that particular window). This research design choice mitigates look-ahead biases and produces predictions that are real time in nature, as our in-sample models mimic the set of information that is publicly available as of the end of each window. This setup is analogous to a regulator or corporate monitor that estimates the likelihood of misreporting at the end of each calendar year using known instances of misreporting over the prior five years plus the disclosure content of annual filings issued over that same five-year period.

The regulator or monitor then uses these estimates to detect misreporting in filings issued in the subsequent year. Panel B of figure 1 illustrates the setup of our prediction analysis for the AA and 10-K/A samples.[21] We cannot apply this setup to the AAER sample, as the CFRM data set does not include the release dates of the AAERs. Thus, the analysis for the AAER sample implicitly assumes that all of the AAER events are publicly known by the end of each estimation window.

Our first prediction model regresses *misreport* on vectors of the disaggregated topic proportions (*topic*) as follows:

$$\log\left(\frac{misreport_{i,t}}{1-misreport_{i,t}}\right)=\alpha+\sum_{j=1}^{31}\beta_j topic_{j,i,t}+\varepsilon_{i,t},\ t\in[T-5,\,T-1],\ i\in\text{Firms}. \quad (1)$$

We estimate equation (1) for the five-year window preceding each of the out-of-sample prediction years in each of the three misreporting samples. As noted previously, our out-of-sample prediction tests cover 1999 to 2010 for the AAER sample and 2005 to 2012 for the AA irregularity sample. We can estimate out-of-sample predictions for all years, 1999 to 2012, in the 10-K/A irregularity sample. Similar to Dechow et al. [2011], we construct a prediction score (*p_misreport*) using the estimated coefficients from equation (1) and apply this scoring in our out-of-sample tests as follows:

$$\log\left(\frac{misreport_{i,T}}{1-misreport_{i,T}}\right)=\alpha+\beta_1 p\_misreport_{i,T}+\varepsilon_{i,T},\ i\in\text{Firms}. \quad (2)$$

We estimate two additional regression specifications to examine RQ1. The first specification replaces the *topic* vector with a vector of quantitative financial measures (denoted *F-score*). The second specification extends equation (1) by including the *topic* and *F-score* vectors as a joint set of predictors.[22] We generate *p_misreport* for both specifications and run the out-of-sample tests. For RQ2, we introduce two additional specifications that include a vector of our textual style measures (denoted *Style*). The first model is a standalone regression of our textual style metrics, whereas the second model incorporates both *topic* and *Style*. We also introduce a comprehensive

[21] As depicted in panel B of figure 1, assume that a firm misstated revenues in 1997. The misstatement was later announced in an amended 10-K filing in 2000 (as captured by our text search tool). Because the misstatement was not revealed until 2000, we exclude the misreporting event from each of the five-year estimation windows ending in 1998 and 1999. The misreporting event would be incorporated only in the 1996–2000 and 1997–2001 estimation windows.

[22] The restructuring indicator variable is valid only for the 2000 fiscal year and onward, due to the lack of restructuring data in Compustat for prior years. We thus exclude the restructuring variable from the *F-Score* vector when estimating the model for the five-year windows that do not overlap with the 2000 fiscal year.

model that includes all three sets of prediction variables: *topic*, *F-score*, and *Style*. We benchmark this model against a final model of *F-score* and *Style* to assess the incremental power of *topic* over the full set of financial metrics and textual features. The comprehensive model is specified in equation 3:

$$\log\left(\frac{misreport_{i,t}}{1-misreport_{i,t}}\right) = \alpha + \sum_{j=1}^{17}\beta_j F\text{-}score_{j,i,t} + \sum_{j=1}^{20}\beta_{j+17} Style_{j,i,t} + \sum_{j=1}^{31}\beta_{j+37} topic_{j,i,t} + \varepsilon_{i,t},\ t \in [T-5, T-1],\ i \in \text{Firms}. \quad (3)$$

We tightly control the convergence of our logistic regressions, given the large number of predictors and the low frequency of misreporting events in our test windows. We control the convergence by conducting checks for both completeness and quasi-completeness of each regression specification. Appendix A.5 of the online appendix details the necessary steps for conducting these checks.

We follow prior research and use the area under the receiver operating characteristics (ROC) curve to evaluate the out-of-sample classification performance of each detection model (Hobson, Mayew, and Venkatachalam [2012], Larcker and Zakolyukina [2012], Bao et al. [2020]).[23] The ROC curve is a two-dimensional plot across different cutoff thresholds of the true positive rate (sensitivity) on the *y*-axis against the false positive rate (specificity) on the *x*-axis. The area under the ROC curve (AUC) is a widely used indicator of a model's predictive ability, because more accurate models would have ROC plots that are closer to the upper left corner of the graph (i.e., higher true positive rates with lower false positive rates). The AUC values can range from 0.50 to 1 and represent the probability that a randomly chosen positive instance of misreporting will be ranked higher by the respective model, compared to a randomly chosen negative instance.

Any reasonable detection model should have an AUC greater than 0.5 (i.e., the model should perform better than a random classification model). We assess whether the AUCs for our models are greater than 0.50 by comparing the AUC for each model against the AUCs produced using simulated random data bootstrapped with 1,000 replications. We also use bootstrapping to compare classification performance across models. We apply the method of Janes, Longton, and Pepe [2009] and conduct bootstrapped,

[23] In robustness tests, we use Fisher's (1932) method to generate an aggregate test statistic that assesses whether one detection model performs better than another when pooled across years. The inferences from this alternative test statistic are consistent with the conclusions drawn from the pooled AUC statistics.

nonparametric Wald tests (based on 1,000 bootstrapped replications) of the differences in the AUCs between the prediction models. Given our rolling time-series analyses, we compute the AUCs and Wald test *p*-values using pooled data for all prediction years with bootstrapped standard errors corrected for clustering by year.

4.2.2. In-Sample Predictive Value of topic. We first evaluate the in-sample performance of *topic* in detecting misreporting. For each sample, we estimate annual in-sample regressions of equation (1) using the unaggregated, industry-normalized topic proportions. For ease of interpretation, figures 2–4 present visual illustrations of the results based on the combined topics listed in appendix A.3 of the online appendix. We exclude the industry-specific topics from the figures for brevity. Figure 2 presents results for the AAER sample; figures 3 and 4 present results for the AA and 10-K/A samples, respectively.

Each figure depicts the presence of each combined topic across the sample's prediction years and whether the combined topic is significantly associated with the misreporting events identified by the respective sample. We report the significance of the combined topics, based on the *z*-statistics for the coefficient estimates on the underlying unaggregated subtopics (i.e., the individual topics associated with a given combined topic in each year). For each prediction year, we present green (red) boxes if at least one subtopic for a given combined topic loads as positive (negative) and significant at the 10% level or greater, and all other subtopics are insignificant. We code the boxes as gray ("Other") if all subtopics for a given combined topic are insignificant, if multiple subtopics are significant but with opposing signs, or if all subtopics are dropped from the regression due to multicollinearity.

We observe in all three figures that the discussion of several topics is relatively consistent across the sample years. These topics include changes in income performance (topics 1 and 2), measurement of postretirement benefits (3), cost commitments (4), and real estate loan operations (9). Other topics appear later in the sample period, indicating the evolving nature of management communications. For instance, discussions of collaborative business arrangements (27, 32, and 37) are prominent in the second half of our prediction period. Likewise, discussions of securitized securities (41) emerge solely in 2008, coinciding with the turmoil in asset-backed securities markets leading up to the financial crisis.

With respect to detection power, we observe in figure 2 (AAER sample) that discussions of increases in income performance, compared to prior periods (combined topic 2), unambiguously predict accounting violations in all but two of the prediction years. Although the direction of the prediction varies throughout the period, we note that discussions of income increases are positively associated with misreporting in 6 out of the 10 years with significant loadings. In the same vein, we find that discussions of segment performance (topic 40) are positively associated with misreporting

FIG. 2.—Combined topic distribution and AAER prediction. This chart depicts the presence of each combined topic across the prediction years and whether the combined topic is significantly associated with financial misreporting involving SEC enforcement actions (AAERs). We do not report the industry-specific topics (topic 50–64) for brevity. The square boxes are color coded based on the direction and statistical significance of the underlying unaggregated subtopics from yearly in-sample logit regressions. The unaggregated subtopics are those topics that are associated with a given combined topic in each year. We orthogonalize all subtopics to two-digit SIC to control for unobserved industry effects. The square boxes are color coded green (red) if the subtopics positively (negatively) predict misreporting involving AAERs. The box is color coded grey if the subtopics are statistically significant with ambiguous direction (i.e., multiple subtopics load in opposing directions), insignificant in the respective year, or dropped from the prediction model due to collinearity.

in at least one prediction year. These results could reflect the upward manipulation of performance measures during periods of misreporting and firms' tendency to grandstand the manipulated performance itself in their disclosures (Dechow et al. [2011], Hoberg and Lewis [2017]). We also observe that discussions of declines in income performance (topic 1) are less

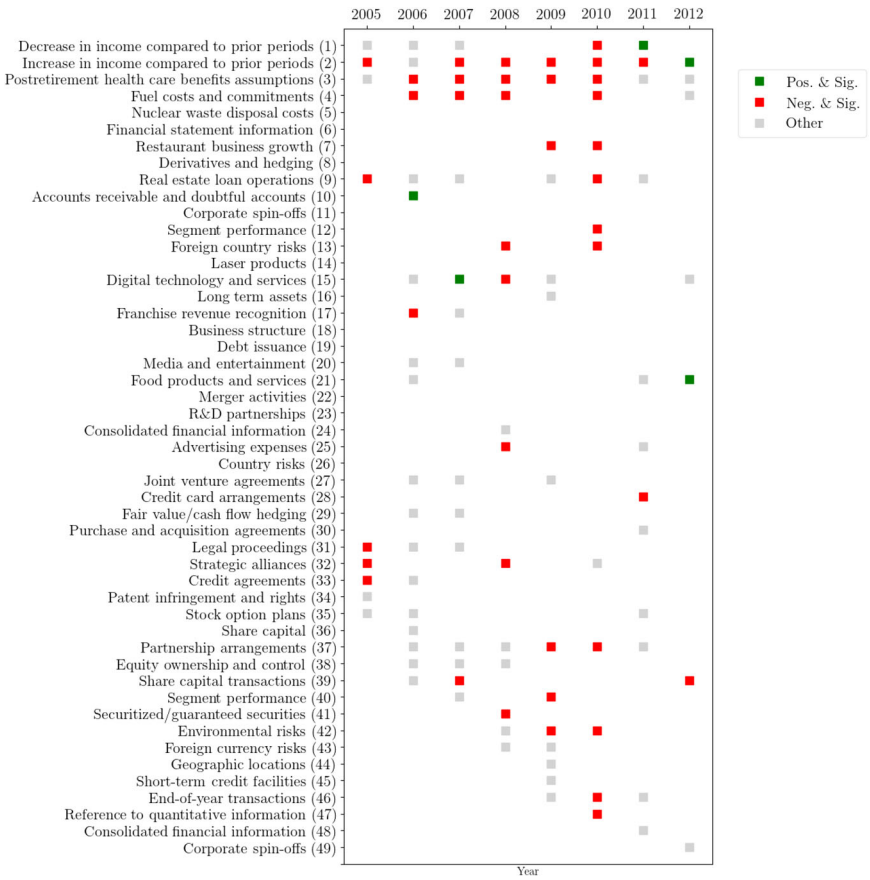


FIG. 3.—Combined topic distribution and AA irregularity prediction. This chart depicts the presence of each combined topic across the prediction period and whether the combined topic is significantly associated with financial misreporting based on the Audit Analytics (AA) irregularity sample. We do not report the industry-specific topics (topic 50–64) for brevity. The square boxes are color coded based on the direction and statistical significance of the underlying unaggregated subtopics from yearly in-sample logit regressions. The unaggregated subtopics are those topics that are associated with a given combined topic in each year. We orthogonalize all subtopics to two-digit SIC to control for unobserved industry effects. The square boxes are color coded green (red) if the subtopics positively (negatively) predict misreporting in the AA sample. The box is color coded grey if the subtopics are statistically significant with ambiguous direction (i.e., multiple subtopics load in opposing directions), insignificant in the respective year, or dropped from the prediction model due to collinearity.

predictive of misreporting in the AAER sample. This finding mirrors prior evidence suggesting that the association between misreporting and poor financial performance is not as clear-cut as posited in the literature (Dechow et al. [2011]).

Other noteworthy results from figure 2 indicate that discussions related to M&As, share transactions, and certain business arrangements

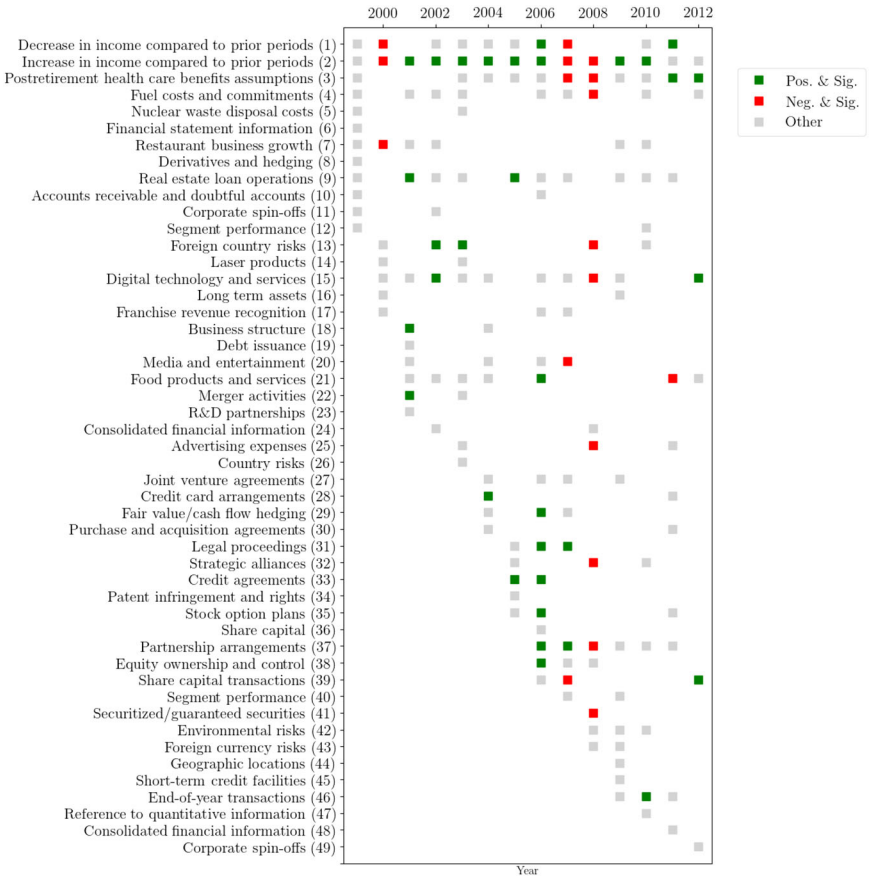


FIG. 4.—Combined topic distribution and 10-K/A irregularity prediction. This chart depicts the presence of each combined topic across the prediction period and whether the combined topic is significantly associated with financial misreporting based on the sample of irregularities identified from amended 10-K filings (10-K/As). We do not report the industry-specific topics (50–64) for brevity. The square boxes are color coded based on the direction and statistical significance of the underlying disaggregated subtopics from yearly in-sample logit regressions. The disaggregated subtopics are those topics that are associated with a given combined topic in each year. We orthogonalize all subtopics to two-digit SIC to control for unobserved industry effects. The square boxes are color coded green (red) if the subtopics positively (negatively) predict misreporting in the 10-K/A sample. The box is color coded grey if the subtopics are statistically significant with ambiguous direction (i.e., multiple subtopics load in opposing directions), insignificant in the respective year, or dropped from the prediction model due to collinearity.

are positively associated with misreporting, whereas discussions surrounding cost commitments, credit agreements, and environmental risks are negatively associated with misreporting. Similar to the findings of Hoberg and Lewis [2017], these results suggest that misreporting firms overdiscuss activities that create incentives to misreport but underdiscuss

issues related to financial agreements and risk factors. We also find that discussions of postretirement benefit assumptions are negatively associated with misreporting in several years. Consistent with Dechow et al. [2011], this result could reflect management's discretionary use of employee benefit assumptions to adjust reported expenses and their tendency to underdiscuss these assumptions in misreporting years.

The evidence from our irregularity samples resembles that from the AAER sample but with variations in the timing and direction of the *topic* loadings. At first glance, the evidence from the AA irregularity sample (figure 3) might seem to contrast with the results from the other samples. But because this sample is confined to the later years of our prediction period, closer inspection reveals that these results are fairly consistent with the post-2005 results in figures 2 and 4. That is, several of the topics that load as positive in the earlier years of the AAER and 10-K/A samples begin to load as negative in the latter half. This evidence further demonstrates the time-varying nature of financial report narratives and how these narratives capture intentional misreporting.

4.2.3. Predictive Value of topic *Versus Financial Variables (RQ1).* Table 2 presents separate summary statistics of our financial variables for misreported and non-misreported firm-years in each sample. We provide tests of differences in the means of each variable (clustered by firm) between each set of firm-years. We find that only one of the financial variables behaves similarly across the three samples. Firms are more likely to issue securities during periods of misreporting in all three samples. Several variables, however, show opposing differences across the samples, with many failing to show significant differences in the AA and 10-K/A irregularity samples. Although these differential results could reflect the more egregious nature of AAERs or identification issues across misreporting samples (KKLM), they highlight the difficulty in establishing clear associations between misreporting and standard financial metrics (Dechow et al. [2011], Purda and Skillicorn [2015]).

Table 3 presents out-of-sample tests of the predictive role of *topic* and *F-score*. We reiterate that these tests are conducted using the *unaggregated* topics (not the combined topics) discovered in each rolling window. Panels A, C, and E present the pooled AUC statistics for the AAER, AA, and 10-K/A samples, respectively. The AUCs across the three panels indicate that quantitative financial metrics (*F-score*) are significant predictors of misreporting, especially for events that trigger an enforcement action. The AUCs for the *F-score* model range from 0.589 in the 10-K/A sample to a high of 0.708 in the AAER sample. All of the AUCs for the *F-score* model are statistically greater than a random classification model (AUC = 0.500) at the 5% level and higher. The AUCs for the *topic* model also exceed the 0.50 threshold in all three samples, indicating the ability of thematic content to independently detect various forms of financial misreporting. The predictive value of *topic* is markedly

TABLE 2
Univariate Statistics for Financial Variables

	Panel A: AAER sample (N = 37, 806)					**Panel B: AA irregularity sample (N = 32, 098)**				
	No AAER ($N = 37, 301$)		AAER ($N = 505$)			No AA Irregularity ($N = 31, 571$)		AA Irregularity ($N = 527$)		
Variable	Mean	Std. Dev.	Mean	Std. Dev.	Diff	Mean	Std. Dev.	Mean	Std. Dev.	Diff
log(*Total Assets*)	5.68	0.0294	6.48	0.152	0.799***	5.89	0.0323	6.13	0.134	0.243**
RSST Accruals	0.0149	0.0012	0.0331	0.0090	0.0182**	0.0105	0.0013	0.0050	0.0093	−0.0055
Δ*Receivables*	0.0093	0.0004	0.0194	0.0033	0.0101***	0.0063	0.0004	0.0111	0.0039	0.0048
Δ*Inventory*	0.0053	0.0003	0.0144	0.0025	0.0091***	0.0042	0.0003	0.0055	0.0023	0.0012
%Soft Assets	0.534	0.0033	0.649	0.0166	0.114***	0.530	0.0037	0.600	0.0160	0.0703***
Δ*Cash Sales*	0.267	0.0664	0.192	0.0187	−0.0757	0.154	0.0889	0.137	0.0248	−0.0172
Δ*Return On Assets*	−0.0030	0.0009	−0.0222	0.0075	−0.0192**	−0.0032	0.0010	0.0036	0.0104	0.0068
Actual Issuance	0.925	0.0024	0.962	0.0128	0.0373***	0.925	0.0027	0.951	0.0119	0.0252**
Operating Leases	0.866	0.0042	0.897	0.0230	0.0315	0.883	0.0044	0.886	0.0218	0.0031
Book-To-Market	0.500	0.0357	0.545	0.0377	0.0447	0.495	0.0424	0.573	0.0910	0.0780
Lag (Mkt-Adj Return)	0.106	0.0046	0.202	0.0393	0.0960**	0.148	0.0050	0.140	0.0359	−0.0084
Merger	0.191	0.0031	0.333	0.0272	0.141***	0.183	0.0034	0.205	0.0194	0.0223
Big N Auditor	0.824	0.0047	0.871	0.0254	0.0468**	0.785	0.0058	0.814	0.0253	0.0295
Mid-size auditor	0.0858	0.0034	0.0653	0.0189	−0.0204	0.107	0.0042	0.0911	0.0190	−0.0156
TotFinancing	0.0431	0.0022	0.0489	0.0094	0.0059	0.0365	0.0024	0.0214	0.0093	−0.0151
ExanteFinancing	−0.0733	0.0106	0.0360	0.0159	0.109***	−0.0537	0.0111	−0.0025	0.0236	0.0512**
Restructuring	0.280	0.0052	0.382	0.0356	0.101***	0.286	0.0052	0.332	0.0281	0.0459

(*Continued*)

TABLE 2—*Continued*

	Panel C: 10-K/A irregularity sample (N = 42,314)				
	No 10-K/A Irregularity ($N = 41{,}617$)		10-K/A Irregularity ($N = 697$)		
Variable	Mean	Std. Dev.	Mean	Std. Dev.	Diff
log(*Total Assets*)	5.76	0.0297	5.74	0.0927	−0.0262
RSST Accruals	0.0153	0.0011	0.0081	0.0105	−0.0072
Δ*Receivables*	0.0091	0.0003	0.0125	0.0025	0.0034
Δ*Inventory*	0.0058	0.0003	0.0029	0.0021	−0.0028
%Soft Assets	0.535	0.0033	0.549	0.0117	0.0143
Δ*Cash Sales*	0.212	0.0698	0.0599	0.123	−0.152
Δ*Return On Assets*	−0.0039	0.0008	−0.0127	0.0099	−0.0088
Actual Issuance	0.923	0.0024	0.941	0.0099	0.0181*
Operating Leases	0.868	0.0041	0.882	0.0149	0.0146
Book-To-Market	0.506	0.0324	0.507	0.0481	0.0010
Lag (Mkt-Adj Return)	0.102	0.0042	0.0667	0.0325	−0.0351
Merger	0.193	0.0031	0.212	0.0164	0.0189
Big N Auditor	0.816	0.0048	0.740	0.0200	−0.0759***
Mid-size auditor	0.0896	0.0035	0.115	0.0142	0.0252**
TotFinancing	0.0408	0.0021	0.0908	0.0138	0.0500***
ExanteFinancing	−0.0613	0.0097	−0.231	0.0500	−0.170***
Restructuring	0.286	0.0052	0.354	0.0211	0.0688***

This table reports summary statistics of our financial statement and stock return variables for misreported and non-misreported firm-years in each sample. Panel A presents summary statistics for the AAER sample, whereas panels B and C present statistics for the Audit Analytics (AA) and 10-K/A irregularity samples, respectively. We conduct two-tailed *t*-tests of the differences in means for each variable between the misreported and non-misreported firm-years in each sample. All standard errors are clustered by firm. Panel A of appendix B provides definitions of each variable. The *Restructuring* variable is valid only for the post-1999 period because restructuring charges were not separately reported in Compustat prior to 2000. The significance levels for the two-tailed *t*-tests are denoted as follows: *** denotes $p < 0.01$, ** denotes $p < 0.05$, and * denotes $p < 0.10$.

TABLE 3
Out-of-Sample Prediction Analysis of topic *and* F-score

Panel A: AUC statistics (AAERs)

Prediction model	AUC
F-score	0.708***
topic	0.680***
topic and *F-score*	0.742***

Panel B: Difference tests (AAERs)

		F-score	*topic* and *F-score*
topic	Diff. in AUC	−0.028	−0.063***
	p-value	(0.19)	(0.00)
topic and *F-score*	Diff. in AUC	0.034***	
	p-value	(0.00)	

Panel C: AUC statistics (AA irregularities)

Prediction model	AUC
F-score	0.597**
topic	0.616**
topic and *F-score*	0.632**

Panel D: Difference tests (AA irregularities)

		F-score	*topic* and *F-score*
topic	Diff. in AUC	0.019	−0.016*
	p-value	(0.29)	(0.10)
topic and *F-score*	Diff. in AUC	0.035***	
	p-value	(0.01)	

Panel E: AUC statistics (10-K/A irregularities)

Prediction model	AUC
F-score	0.589***
topic	0.616***
topic and *F-score*	0.630***

Panel F: Difference tests (10-K/A irregularities)

		F-score	*topic* and *F-score*
topic	Diff. in AUC	0.027**	−0.014*
	p-value	(0.08)	(0.08)
topic and *F-score*	Diff. in AUC	0.041***	
	p-value	(0.00)	

This table reports results from comparative out-of-sample tests of the prediction models based on vectors of *topic* and financial metrics (denoted as *F-score*). The *topic* vector is comprised of measures capturing the proportion of a firm's 10-K filing devoted to discussing a particular theme. Section 3.2.2 describes how the topic measures are derived. The set of variables in the *F-score* vector are defined in panel A of appendix B. Panels A, C, and E present statistics of the detection performance of each model for the AAER sample and the Audit Analytics (AA) and 10-K/A irregularity samples, respectively. The performance statistics are computed using pooled values of the area under the receiver operating classification curve (AUC) generated by out-of-sample regressions of *misreport* on *p_misreport* (derived by estimating each model in rolling five-year windows prior to the prediction year). The statistical significance of each AUC statistic is determined by assessing whether the statistic is significantly greater than 0.50, which is the AUC for a random classification model. We determine the statistical significance by comparing the statistic against the AUCs produced using simulated random data bootstrapped and clustered by year with 1,000 replications. Panels B, D, and F present nonparametric Wald tests of the differences in the AUCs for the AAER, AA, and 10-K/A samples, respectively. The Wald tests are based on 1,000 bootstrap iterations with clustering by year. Each panel reports test statistics and *p*-values (in parentheses) indicating whether the model specification in a given row is significantly better at predicting misreporting out-of-sample compared to the model specification in the respective column (two-tailed). The significance levels for all tests are denoted as follows: *** denotes $p < 0.01$, ** denotes $p < 0.05$, and * denotes $p < 0.10$.

higher in the AAER sample, with a predictive gain of 18% over a random model, compared to a 12% gain in the AA and 10-K/A samples (AUC of 0.680 in the AAER sample versus 0.616 for both irregularity samples; statistically different at the 1% and 5% levels). This differential result suggests that disclosure content varies strongly with instances of misreporting where there is high confidence of the intent to mislead.

Panels B, D, and F present Wald tests of the predictive ability of *topic*, compared to that of *F-score* and a joint model of both predictors. Panel B presents the test results for the AAER sample, whereas panels D and F present results for the AA and 10-K/A samples, respectively. Each panel compares the model listed in the row with the model listed in the column. The AUC comparisons for the AAER sample (panel B) reveal that the standalone *topic* vector performs equivalent to the *F-score* vector. Specifically, the pooled AUC for the standalone *topic* model is not significantly different from the AUC for the *F-score* model (Wald *p*-value = 0.19). Thus, *topic* as a standalone predictor rivals the detection ability of quantitative financial metrics. Nonetheless, we find that a joint model of *topic* and *F-score* performs significantly better at predicting AAERs, compared to the standalone *F-score* model. In fact, our *topic* measure increases predictive accuracy by 3.4% when added to the *F-score* model (AUC of 0.742 versus 0.708, Wald *p-value* = 0.00). The joint *topic* and *F-score* model also outperforms the standalone *topic* model by 6.2% (AUC of 0.742 versus 0.680, Wald *p*-value = 0.00), consistent with our finding that *F-score* is an equally strong predictor of AAER events.

The benchmarking results from our irregularity samples continue to demonstrate the incremental predictive value of disclosure content. In the AA sample (panel D), we find that *topic* improves predictive accuracy by 3.5% (AUC of 0.632 versus 0.597, Wald *p*-value = 0.00) when added to the standard *F-score* model. Similar to the AAER sample, we find no difference in the predictive accuracy of the standalone *topic* and *F-score* models for the AA sample. That is, the disclosure content of annual reports fares just as well in detecting irregularity restatements, compared to quantitative financial metrics. Lastly, the 10-K/A results in panel F show that our *topic* measure boosts predictive accuracy by a magnitude similar to that observed in the AAER and AA samples (4.1% increase in AUC, Wald *p*-value = 0.00).

Collectively, the results in table 3 suggest that content-based information drawn from annual report narratives improves the detection of misreporting beyond what can be achieved by financial metrics. Our evidence also suggests that disclosure content and financial variables performs equally well in detecting misreporting, though both predictors serve as complementary warning signals.

4.2.4. The Predictive Value of topic *Versus Textual Style (RQ2).* Table 4 presents separate summary statistics for our style characteristics. We find that many of the style features shift inconsistently in misreporting years and, at times, contradict conventional views. For instance, misreported

TABLE 4
Univariate Statistics for Textual Style Variables

	Panel A: AAER sample ($N = 37,806$)					Panel B: AA irregularity sample ($N = 32,098$)				
	No AAER ($N = 37,301$)		AAER ($N = 505$)			No AA Irregularity ($N = 31,571$)		AA Irregularity ($N = 527$)		
Variable	Mean	Std. Dev.	Mean	Std. Dev.	Diff	Mean	Std. Dev.	Mean	Std. Dev.	Diff
Length										
Parsed Size	171687	1257	167777	6439	−3910	201755	1452	197514	6674	−4240
Sentence Length	23.8	0.0251	23.5	0.173	−0.259	24.2	0.0254	24.0	0.131	−0.194
Complexity										
Word Stddev	3.07	0.0009	3.08	0.0055	0.0076	3.07	0.0009	3.07	0.0048	−0.0030
Paragraph Stddev	4.06	0.0375	3.81	0.214	−0.248	5.84	0.0961	5.13	0.343	−0.709**
Variation										
Repetitions	0.0794	0.0005	0.0831	0.0035	0.0037	0.0859	0.0006	0.0945	0.0032	0.0086***
Sentence Stddev	16.5	0.0451	16.5	0.302	−0.0401	15.8	0.0386	16.1	0.221	0.292
Type Token Ratio	0.126	0.0006	0.126	0.0042	−0.0002	0.112	0.0006	0.112	0.0026	−0.0007
Readability										
Coleman-Liau Index	14.5	0.0087	14.6	0.0520	0.0938*	14.5	0.0087	14.5	0.0473	−0.0011
Fog	17.8	0.0161	17.7	0.117	−0.0848	18.3	0.0151	18.0	0.0768	−0.227***
Tense										
% Active Voice	0.606	0.0007	0.599	0.0052	−0.0074	0.625	0.0006	0.614	0.0031	−0.0111***
% Passive Voice	0.0314	0.0003	0.0308	0.0011	−0.0006	0.0314	0.0003	0.0321	0.0011	0.0007
Word Choice										
% Negative	0.0126	0.0001	0.0129	0.0004	0.0003	0.0141	0.0001	0.0139	0.0004	−0.0002
% Positive	0.0064	0.0000	0.0062	0.0001	−0.0002	0.0067	0.0000	0.0066	0.0001	−0.0001

(*Continued*)

TABLE 4—*Continued*

	Panel A: AAER sample ($N = 37,806$)					Panel B: AA irregularity sample ($N = 32,098$)				
	No AAER ($N = 37,301$)		AAER ($N = 505$)			No AA Irregularity ($N = 31,571$)		AA Irregularity ($N = 527$)		
Variable	Mean	Std. Dev.	Mean	Std. Dev.	Diff	Mean	Std. Dev.	Mean	Std. Dev.	Diff
Emphasis										
All Caps	567	6.03	582	30.5	15.2	564	6.87	596	35.5	32.1
Exclamation Points	0.371	0.100	0.0792	0.0292	−0.292***	0.360	0.0933	0.319	0.203	−0.0411
Question Marks	0.0931	0.0348	0.0673	0.0326	−0.0258	0.112	0.0483	0.0493	0.0231	−0.0628
Processing										
log (*Bullets*)	5.22	0.0232	5.03	0.136	−0.195	5.27	0.0263	5.20	0.132	−0.0666
Header	1,452	5.01	1,445	11.6	−6.61	1,456	5.47	1,460	8.17	4.60
Newlines	1,590	16.8	1,266	72.3	−325***	2,062	21.2	1,824	114	−238**
Tags	471,954	6,524	325,174	33,883	−146,780***	738,661	9,033	617,963	45,660	−120,697***

	Panel C: 10-K/A irregularity sample ($N = 42,314$)				
	No 10-K/A Irregularity ($N = 41,617$)		10-K/A Irregularity ($N = 697$)		
Variable	Mean	Std. Dev.	Mean	Std. Dev.	Diff
Length					
Parsed Size	177,540	1,279	237,193	5,246	59,653***
Sentence Length	23.9	0.0247	24.6	0.0962	0.626***
Complexity					
Word Stddev	3.07	0.0009	3.07	0.0028	0.0032
Paragraph Stddev	5.12	0.0729	6.92	0.608	1.80***

(*Continued*)

TABLE 4—*Continued*

	Panel C: 10-K/A irregularity sample ($N = 42,314$)				
	No 10-K/A Irregularity ($N = 41,617$)		10-K/A Irregularity ($N = 697$)		
Variable	Mean	Std. Dev.	Mean	Std. Dev.	Diff
Variation					
Repetitions	0.0791	0.0005	0.0994	0.0026	0.0203***
Sentence Stddev	16.4	0.0416	16.3	0.167	−0.117
Type Token Ratio	0.124	0.0006	0.101	0.0016	−0.0237***
Readability					
Coleman-Liau Index	14.5	0.0084	14.4	0.0249	−0.0662***
Fog	17.9	0.0159	18.4	0.0535	0.501***
Tense					
% Active Voice	0.611	0.0007	0.625	0.0025	0.0142***
% Passive Voice	0.0313	0.0003	0.0312	0.0008	−0.0000
Word Choice					
% Negative	0.0129	0.0001	0.0161	0.0002	0.0032***
% Positive	0.0064	0.0000	0.0065	0.0001	0.0001
Emphasis					
All Caps	556	5.88	702	42.2	146***
Exclamation Points	0.354	0.0899	0.475	0.124	0.121
Question Marks	0.0890	0.0358	0.347	0.312	0.258
Processing					
log (*Bullets*)	5.19	0.0227	5.47	0.0890	0.280***
Header	1,429	4.73	1,475	10.4	45.9***
Newlines	1,698	17.4	2,144	82.5	446***
Tags	555,212	7,411	793,039	37,437	237,827***

This table reports summary statistics of our textual style variables for misreported and non-misreported firm-years in each sample. Panel A presents summary statistics for the AAER sample, whereas panels B and C present statistics for the Audit Analytics (AA) and 10-K/A irregularity samples, respectively. We conduct two-tailed *t*-tests of the differences in means for each variable between the misreported and non-misreported firm-years in each sample. All standard errors are clustered by firm. Panel B of appendix B provides definitions of each variable. The significance levels for the two-tailed *t*-tests are denoted as follows: *** denotes $p < 0.01$, ** denotes $p < 0.05$, and * denotes $p < 0.10$.

filings in the AA and 10-K/A samples are more readable, relative to nonmisreported filings, whereas misreported filings in the AAER sample are less readable. Such opposing evidence is not unique to our study, as Purda and Skillicorn [2015] find contradictory evidence of more deceptive and negative language in filings classified as "truthful." These findings underscore the potential pitfalls in relying on basic textual measures to identify financial misreporting.

Table 5 presents the pooled AUC statistics for out-of-sample tests of the predictive performance of *topic*, relative to the performance of the textual style (*Style*) vector. Panels A and B present the test statistics for AAERs; panels C–F present the results for the two irregularity samples. The benchmarking tests for the AAER and AA samples indicate that *topic* by itself better predicts misreporting than the standalone *Style* model (3.1% and 3.5% increase in AUC, in panels B and D, respectively; Wald *p*-values = 0.01 and 0.06). The *topic* model, however, performs worse than the *Style* specification in the 10-K/A sample (panel F).

We observe that a joint model of *topic* and *Style* outperforms the individual *Style* model by almost 6% in the AAER sample (panel B, Wald *p*-value = 0.00). Both models perform at the same level in the AA and 10-K/A samples (panels D and F, respectively; Wald *p*-values = 0.13 and 0.27). Although the joint *topic* and *Style* model dominates in the AAER sample, the basic *topic* model achieves an accuracy level in the AA sample that ranks just as high as the joint model. Thus, the predictive value of *topic* over *Style* is quite strong when detecting misreporting events that trigger an enforcement action or a financial restatement.

4.2.5. Joint Predictive Value of topic*, Financial Variables, and Textual Style.* We next examine the interplay between all three sets of predictors: *topic*, *F-Score*, and *Style*. This comprehensive analysis evaluates whether the incremental predictive ability of *topic* is robust to the inclusion of the combined set of financial and textual style measures. Table 6 presents out-of-sample results for equation (3). For brevity, we do not present performance measures for all model combinations of the three sets of predictors. For the AAER and AA samples, we present performance statistics for the joint *topic* and *F-score* model, because the AUCs for this specification outrank the AUCs for the joint *topic* and *Style* model in these two samples. Likewise, we present performance statistics for the joint *topic* and *Style* model for the 10-K/A sample, because this model is more dominant in that sample.[24]

In table 6, we find that the three-vector model performs well in detecting misreporting out of sample. The AUCs across the three samples are well

[24] The AUCs for the joint *topic* and *F-score* model are 3.4% and 1.8% higher than the AUCs for the joint *topic* and *Style* model in the AAER and AA samples, respectively. The AUC for the *topic* and *Style* specification is 3.9% higher than that for the *topic* and *F-score* model in the 10-K/A sample. We compute these differences by comparing the AUCs reported in tables 3 and 5 across each sample. All the differences are statistically significant at the 5% level and greater, except for the 1.8% difference quoted for the AA sample.

TABLE 5
Out-of-Sample Prediction Analysis of topic *and* Style

Panel A: AUC statistics (AAERs)

Prediction model	AUC
Style	0.649***
topic	0.680***
topic and *Style*	0.708***

Panel B: Pooled ROC AUC difference tests (AAERs)

		Style	*topic* and *Style*
topic	Diff. in AUC	0.031***	−0.028***
	p-value	(0.01)	(0.00)
topic and *Style*	Diff. in AUC	0.059***	
	p-value	(0.00)	

Panel C: AUC statistics (AA irregularities)

Prediction model	AUC
Style	0.581**
topic	0.616**
topic and *Style*	0.614**

Panel D: Pooled ROC AUC difference tests (AA irregularities)

		Style	*topic* and *Style*
topic	Diff. in AUC	0.035*	0.002
	p-value	(0.06)	(0.85)
topic and *Style*	Diff. in AUC	0.033	
	p-value	(0.13)	

Panel E: AUC statistics (10-K/A irregularities)

Prediction model	AUC
Style	0.663***
topic	0.616***
topic and *Style*	0.669***

Panel F: Pooled ROC AUC difference tests (10-K/A irregularities)

		Style	*topic* and *Style*
topic	Diff. in AUC	−0.047***	−0.054***
	p-value	(0.00)	(0.00)
topic and *Style*	Diff. in AUC	0.007	
	p-value	(0.27)	

This table reports results from comparative out-of-sample tests of the prediction models based on vectors of *topic* and textual style metrics (denoted as *Style*). The *topic* vector is comprised of measures capturing the proportion of a firm's 10-K filing devoted to discussing a particular theme. Section 3.2.2 describes how the topic proportions are derived. The set of variables in the *Style* vector are defined in panel B of appendix B. Panels A, C, and E present statistics of the detection performance of each model for the AAER sample and the Audit Analytics (AA) and 10-K/A irregularity samples, respectively. The performance statistics are computed using pooled values of the area under the receiver operating classification curve (AUC) generated by out-of-sample regressions of *misreport* on *p_misreport* (derived by estimating each model in rolling five-year windows prior to the prediction year). The statistical significance of each AUC statistic is determined by assessing whether the statistic is significantly greater than 0.50 (the AUC for a random classification model). The statistical significance is determined by comparing the statistic against the AUCs produced using simulated random data bootstrapped and clustered by year with 1,000 replications. Panels B, D, and F present non-parametric Wald tests of the differences in the AUCs for the AAER, AA, and 10-K/A samples, respectively. The Wald tests are based on 1,000 bootstrap iterations with clustering by year. Each panel reports test statistics and *p*-values (in parentheses) indicating whether the model specification in a given row is significantly better at predicting misstatements out-of-sample compared to the model specification in the respective column (two-tailed). The significance levels for all tests are denoted as follows: *** denotes $p < 0.01$, ** denotes $p < 0.05$, and * denotes $p < 0.10$.

TABLE 6
Out-of-Sample Prediction Analysis of topic*,* F-score*, and* Style

Panel A: AUC statistics (AAERs)

Prediction model	AUC
F-score and *Style*	0.719***
topic, *F-score*, and *Style*	0.752***
topic	0.680***
topic and *F-score*	0.742***

Panel B: Pooled ROC AUC difference tests (AAERs)

		F-score and *Style*	*topic*, *F-score*, and *Style*
topic	Diff. in AUC	−0.040**	−0.072***
	p-value	(0.04)	(0.00)
topic and *F-score*	Diff. in AUC	0.023***	−0.009
	p-value	(0.01)	(0.13)
topic, *F-score*, and *Style*	Diff. in AUC	0.032***	
	p-value	(0.00)	

Panel C: AUC statistics (AA irregularities)

Prediction model	AUC
F-score and *Style*	0.606***
topic, *F-score*, and *Style*	0.635***
topic	0.610***
topic and *F-score*	0.632***

Panel D: Pooled ROC AUC difference tests (AA irregularities)

		F-score and *Style*	*topic*, *F-score*, and *Style*
topic	Diff. in AUC	0.010	−0.018
	p-value	(0.45)	(0.14)
topic and *F-score*	Diff. in AUC	0.026**	−0.003
	p-value	(0.04)	(0.80)
topic, *F-score*, and *Style*	Diff. in AUC	0.028***	
	p-value	(0.01)	

Panel E: AUC statistics (10-K/A irregularities>)

Prediction model	AUC
F-score and *Style*	0.667***
topic, *F-score*, and *Style*	0.670***
topic	0.616***
topic and *Style*	0.669***

(*Continued*)

TABLE 6—*Continued*

Panel F: Pooled ROC AUC difference tests (10-K/A irregularities)

		F-score and *Style*	*topic*, *F-score*, and *Style*
topic	Diff. in AUC	−0.051***	−0.054***
	p-value	(0.00)	(0.00)
topic and *Style*	Diff. in AUC	0.003	−0.001
	p-value	(0.59)	(0.74)
topic, *F-score*, and *Style*	Diff. in AUC	0.004	
	p-value	(0.42)	

This table reports results from comparative out-of-sample tests of the prediction models based on vectors of *topic*, quantitative financial metrics (*F-score*), and textual style feature (*Style*). The *topic* vector is comprised of measures capturing the proportion of a firm's 10-K filing devoted to discussing a particular theme. Section 3.2.2 describes how the topic measures are derived. The set of variables in the *F-score* and *Style* vectors are defined in appendix B. Panels A, C, and E present statistics of the detection performance of each model for the AAER sample and the Audit Analytics (AA) and 10-K/A irregularity samples, respectively. The performance statistics are computed using pooled values of the area under the receiver operating classification curve (AUC) generated by out-of-sample regressions of *misreport* on *p_misreport* (derived by estimating each model in rolling five-year windows prior to the prediction year). The statistical significance of each AUC statistic is determined by assessing whether the statistic is significantly greater than 0.50 (the AUC for a random classification model). The statistical significance is determined by comparing the statistic against the AUCs produced using simulated random data bootstrapped and clustered by year with 1,000 replications. Panels B, D, and F present nonparametric Wald tests of the differences in the AUCs for the AAER, AA, and 10-K/A samples, respectively. The Wald tests are based on 1,000 bootstrap iterations with clustering by year. Each panel reports test statistics and *p*-values (in parentheses) indicating whether the model specification in a given row is significantly better at predicting misstatements out-of-sample compared to the model specification in the respective column (two-tailed). The significance levels for all tests are denoted as follows: *** denotes $p < 0.01$, ** denotes $p < 0.05$, and * denotes $p < 0.10$.

above the 0.50 threshold, ranging from 0.635 in the AA sample (panel C) to as high as 0.752 in the AAER sample (panel A). The benchmarking tests indicate that the addition of *topic* to the *F-score* and *Style* model improves predictive accuracy by 3.2% for AAERs (Wald *p*-value = 0.00 in panel B) and 2.8% for irregularity restatements (Wald *p*-value = 0.01 in panel D). The incremental value of *topic* is insignificant in the 10-K/A sample when we compare the three-vector model to the joint *F-score* and *Style* model. We also find that the three-vector model does not perform any better than the joint *topic* and *F-score* model in the AAER and AA samples (Wald *p*-value = 0.13 and 0.80 in panels B and D) or the *topic* and *Style* model in the 10-K/A sample (Wald *p*-value = 0.74 in panel F). This evidence corroborates our previous results: *topic* and *F-Score* are strong predictors of AAERs and restatements, whereas *topic* and *Style* provide robust power for detecting broad misrepresentations and disclosure omissions.

4.3 THE ECONOMIC SIGNIFICANCE OF *topic*

We gauge the economic significance of our topic measure by examining the out-of-sample classification accuracy of our detection models at the 50th, 90th, and 95th percentiles of the predicted probability scores. Following Dechow et al. [2011], we consider scores above the 50th percentile as "above normal risk" and those above the 90th and 95th percentiles as "high risk." The accuracy rates for each prediction model are equivalent to the true positive rate or sensitivity of the model at the various cutoffs. We

compute these rates for each prediction year and report the average annual percentage of misreported filings that are accurately classified by each model at the respective percentile cutoff. We also report the total number of misreported filings that are correctly classified at each cutoff.

In addition, we follow Bao et al. [2020] and use the Normalized Discounted Cumulative Gain at the position k (NDCG@k) as an alternative measure of classification accuracy, where k is the top 1% or 99th percentile of the predicted probability scores. The NDCG@k measure uses a logarithmic function to weight the ranks of the predicted scores from a given model, such that the top prediction scores have higher values than lower prediction scores. False positives receive a rank of zero. These weights are then summed to arrive at the Discounted Cumulative Gain at the cutoff percentile k (denoted DCG@k). We set k at 1%, so that the predicted scores in the ranking list represent the 99th percentile of the test sample in a given prediction year. The DCG@k measure is further normalized to values ranging from 0 to 1, to allow for comparisons of ranking quality across multiple models.

Table 7 reports the percentage and number of filings that are correctly classified as misreported using multiple models. Panel A reports results for the AAER sample; panels B and C report results for the AA and 10-K/A samples, respectively. For the AAER sample, we find that the basic *F-score* model correctly classifies 72.51% of misreported filings at the 50th percentile cutoff (total count of 319 filings). The joint model of *topic* and *F-Score* or, alternatively, the three-vector model performs better, flagging about 79% of misreported filings on average. When we focus on high-risk prediction scores, we observe that the three-vector model captures the most misreporting events at the 90th and 95th percentiles (32.90% and 22.44% or total counts of 150 and 96 misreported filings, respectively).

To quantify the economic value of *topic*, we note that the classification rate at the 95th percentile improves by 59% when *topic* is added to the benchmark *F-score* and *Style* model (accuracy rate of 22.44% versus 14.11%).[25] In terms of raw numbers, we capture 26 additional misreporting firms at the 95th percentile when *topic* is added to the benchmark model. This relative improvement is striking when we consider the low frequency of AAER filings and the high costs associated with misreporting events that are not detected by traditional prediction models (Beneish and Vorst [2019]).[26] The classification rate at the 90th percentile also increases by 36% when *topic* is added to the benchmark model (32.90% versus 24.28%).

[25] We determine the relative improvement in percentage terms by taking the change in the classification accuracy rate and dividing through by the classification rate of the benchmark model, that is, $[(22.44 - 14.11)/14.11 = 59\%]$. We use the same approach when assessing incremental value based on the NDCG@k statistic.

[26] Beneish and Vorst [2019] estimate that investors suffer three-day return losses of $447 million for misreporting events that are not detected by models built using traditional financial statement variables.

TABLE 7
Out-of-Sample Classification Performance of topic

Panel A: Classification of AAERs

	50th Percentile		90th Percentile		95th Percentile		NDCG@k
Prediction model	%	Count	%	Count	%	Count	99th
topic	70.29	292	22.75	91	10.39	47	0.113
F-score	72.51	319	22.95	111	11.24	65	0.146
Style	65.73	273	14.60	62	6.78	35	0.079
topic and *F-score*	78.56	332	28.93	140	18.01	86	0.176
topic and *Style*	74.79	312	21.40	91	13.38	52	0.118
F-score and *Style*	76.42	329	24.28	117	14.11	70	0.163
topic, *F-score*, and *Style*	79.43	339	32.90	150	22.44	96	0.188

Panel B: Classification of AA irregularities

	50th Percentile		90th Percentile		95th Percentile		NDCG@k
Prediction model	%	Count	%	Count	%	Count	99th
topic	57.36	142	17.07	39	8.68	18	0.062
F-score	60.84	140	11.03	28	4.63	14	0.086
Style	55.67	136	15.04	33	7.79	19	0.096
topic and *F-score*	61.41	151	17.28	41	8.16	20	0.106
topic and *Style*	56.81	144	14.61	38	9.71	24	0.132
F-score and *Style*	53.26	132	12.05	29	6.44	17	0.106
topic, *F-score*, and *Style*	62.21	156	14.07	35	9.68	24	0.124

Panel C: Classification of 10-K/A irregularities

	50th Percentile		90th Percentile		95th Percentile		NDCG@k
Prediction model	%	Count	%	Count	%	Count	99th
topic	65.19	419	19.74	117	10.49	65	0.113
F-score	58.33	368	15.96	104	9.95	63	0.141
Style	69.68	453	25.54	165	14.06	91	0.079
topic and *F-score*	63.31	404	21.40	133	11.30	74	0.176
topic and *Style*	70.61	457	24.55	160	15.56	100	0.118
F-score and *Style*	69.60	451	26.72	174	15.94	98	0.172
topic, *F-score*, and *Style*	70.18	455	28.05	179	16.83	105	0.180

This table reports the classification accuracy of our prediction models using out-of-sample prediction scores. Panel A reports the results for the AAER sample, whereas panels B and C present results for the Audit Analytics (AA) and 10-K/A irregularity samples, respectively. In each panel, we report the average annual percentage (%) and total number (Count) of misreported 10-K filings that are accurately classified as misreported by each prediction model at the respective cutoffs at the 50th, 90th, and 95th percentiles of the predicted probability scores. The final column of each panel presents the NDCG@k score for each prediction model, where k is the 99th percentile or the top 1% of the predicted scores. The NDCG@k measure evaluates the ranking quality of each prediction model and ranges from 0 to 1, with higher values indicating greater classification performance.

This increase leads to 33 more filings being correctly classified at this cutoff. These findings corroborate our inference that *topic* is incrementally valuable in detecting misreporting. The same interpretation holds when we evaluate the detection rates at the 99th percentile based on the NDCG@k measure. Here, the inclusion of *topic* in the three-vector model improves detection accuracy by 15%, compared to the benchmark *F-score* and *Style* model (0.188 versus 0.163).

Our results for the AA sample (panel B) further demonstrate the economic value of our topic measure. We find that *topic* increases detection accuracy by 50% at the 95th percentile cutoff when added to the benchmark model (classification accuracy of 9.68% versus 6.44%). We also observe that *topic* by itself or joint models of *topic* paired with *F-score* or *Style* perform just as well or higher than the three-vector model at the 90th and 95th percentile cutoffs. Thus, there is little value in adding *F-score* or *Style* as a third predictor of high-risk restatements, once *topic* is included in the model. In panel C, the three-vector model is most efficient at detecting high-risk events in the 10-K/A sample. For instance, the inclusion of *topic* in the three-vector model improves classification accuracy by roughly 6% at the 95th percentile cutoff when benchmarked against the joint *F-score* and *Style* model (16.83% versus 15.94%). This improvement equates to an additional seven filings being correctly classified as misreported. Collectively, the results in table 7 suggest that the incremental predictive power of *topic* is economically significant and that its value is quite salient when detecting high-risk reporting practices.

Following prior studies (e.g., Dechow et al. [2011]), we further illustrate the economic significance of our results by assessing the incremental value of *topic* in detecting the Enron accounting scandal. We focus on the prediction scores from the three-vector model to ensure that all possible predictors are considered. The earliest out-of-sample year in our analysis is 1999. Thus, we restrict our analysis to the 10-K filings issued by Enron in 1999 and 2000, that is, the misreported filings for fiscal 1998 and 1999, respectively.[27]

The three-vector model classifies Enron's 1999 filing as misreported based on a prediction score that ranks at the 93rd percentile across all filings issued in 1999. Two variables contribute the most to Enron's prediction score. The first is firm size (log of total assets), consistent with the notion that large firms are more likely to attract SEC scrutiny (Files [2012]). The second variable is the proportion of the 10-K filing devoted to discussing year-over-year increases in income (combined topic 2). Interestingly, Enron's industry-normalized value for this topic proportion ranks at the 98th percentile. We further find that the 10-K filing issued by Enron in 2000 is classified as misreported at an even higher percentile (98.5). Enron's discussion of income increases is again the biggest contributor to its

[27] The enforcement actions against Enron cited material accounting violations for the 1997 to 2000 fiscal years. See SEC AAER Numbers 1640 and 1821.

prediction score. But this time, the topic proportion is substantially lower in the 2000 filing, ranking in just the second percentile, relative to industry peers. This dramatic drop in the attention devoted to this topic could reflect deliberate efforts by Enron executives to distract attention from soaring (manipulated) earnings and the sources of its revenue growth.[28,29]

5. *Extended Analyses and Robustness Tests*

5.1 CONTROLLING FOR "REPEAT OFFENDERS"

Our rolling window procedure requires a misreporting event to be publicly known by the end of the estimation window for the event to be included in our in-sample estimations. (See earlier discussion in section 4.2.1.) There are no cases in our samples where the same misreporting event affects multiple annual reports spanning both the in-sample estimation window and the out-of-sample prediction year. Thus, the only way for a misreporting firm to be included in both periods is for the firm to have two or more unrelated misreporting events (i.e., at least one event appearing in the estimation window and another separate event appearing in the prediction year). Although such cases are due to unrelated misreporting events, their inclusion raises the concern that our *topic* measure could be biased toward identifying repeat offenders or certain types of firms, rather than detecting variations in thematic content when firms misreport.

We alleviate this concern by imposing an additional sample restriction to ensure that our prediction periods are closer to being out-of-sample with respect to a given firm. Specifically, we remove misreporting firms from the out-of-sample prediction period if *misreport* is set to 1 in any year during the in-sample estimation window. That is, repeat offenders are retained in the estimation period in a given rolling five-year window but excluded from the sample in the prediction year. Note that this sample restriction only affects the observations that appear in our out-of-sample tests; the observations in the estimation windows are unchanged because repeat-offender firms are retained in-sample.[30]

Table 8 reports the frequency of misreporting for our prediction years after removing repeat-offender firms. The sample adjustment is quite restrictive as we are left with a significantly smaller number of misreporting events in the out-of-sample periods. We lose 64% of the AAER events across our

---

[28] In a March 2001 *Fortune* article (McLean [2001]), then-CFO Andrew Fastow noted "competitive reasons" when explaining Enron's suppression of income sources in its financial reports.

[29] We conduct a second case study of the AAER filed against Zale Corporation for the improper capitalization of television advertising costs between 2004 and 2009 (SEC AAER No. 3270). The three-vector model correctly classifies Zales' 10-K filings as misreported at the 97th percentile and above in all years except 2004. The topics that contribute the most to Zales' prediction score point to extensive discussion related to media and entertainment (combined topic 20) and digital technology and services (combined topic 15). A manual review of the

TABLE 8

Distribution of Financial Misreporting for Out-of-Sample Prediction Periods: Controlling for Repeat Offenders

	Panel A: AAERs			Panel B: AA irregularities			Panel C: 10-K/A irregularities		
Year	Observations	Frequency	Percentage	Observations	Frequency	Percentage	Observations	Frequency	Percentage
1999	2,195	21	0.96				2,195	11	0.50
2000	2,041	22	1.08				2,041	20	0.98
2001	2,021	16	0.79				2,021	16	0.79
2002	2,391	20	0.84				2,391	24	1.00
2003	2,936	23	0.78				2,936	49	1.67
2004	2,843	8	0.28				2,843	61	2.15
2005	2,678	4	0.15	2,678	19	0.71	2,678	53	1.98
2006	2,608	7	0.27	2,608	15	0.58	2,608	62	2.38
2007	2,549	2	0.08	2,549	12	0.47	2,549	44	1.73
2008	2,535	3	0.12	2,535	7	0.28	2,535	35	1.38
2009	2,564	6	0.23	2,564	16	0.62	2,564	47	1.83
2010	2,424	2	0.08	2,424	10	0.41	2,424	34	1.40
2011				2,330	7	0.30	2,330	32	1.37
2012				2,178	6	0.28	2,178	30	1.38
Prediction firm-years	29,785	134	0.45	19,866	92	0.46	34,293	518	1.51
No. of firms	5,259	133		3,916	91		5,427	511	

This table reports the frequency and percentage of financial misreporting events for the out-of-sample prediction years in each sample after removing misreporting events by repeat-offender firms. We classify a firm as a repeat offender if the firm has a misreporting event in both the estimation window and the respective prediction year. For such cases, we remove the misreporting event from the prediction year, while retaining the events in the estimation window. Panel A presents the yearly frequencies for the AAER sample, whereas panels B and C present the frequencies for the Audit Analytics (AA) and 10-K/A irregularity samples, respectively. Each sample excludes firm observations with missing data in Compustat and CRSP to compute the set of quantitative financial statement and stock return variables used in our prediction models. The final rows of each panel report overall frequencies and percentages for the out-of-sample prediction periods. The prediction years for the AAER sample (panel A) extends from 1999 through 2010 and excludes the 2011 and 2012 years due to the lack of misreporting events in those two years. The prediction years for the AA sample (panel B) extends from 2005 through 2012, because the AA restatement database is incomplete prior to 2000. The predictions years for the 10-K/A sample (panel C) extend from 1999 through 2012. The first five-year estimation window for the AAER and 10-K/A samples runs from 1994 through 1998, whereas the first estimation window for the AA sample runs from 2000 through 2004.

prediction years, resulting in a misreporting rate of 0.45% compared to the 1.41% reported in table 1 (a drop from 419 to 134 events for the 1999–2010 period). We also lose 61% of the events in the AA sample (from 234 to 92 events) and about 21% in the 10-K/A sample (from 648 to 518 events). The out-of-sample misreporting rates for these two samples are also dramatically lower at 0.46% and 1.51%, respectively. Thus, although the repeat-offender restriction helps to alleviate identification concerns, it exacerbates the relative rarity of misreporting and, as a consequence, may lead to noisy model predictions (Perols et al. [2017]).

We report replicated results for our economic significance tests in panels A–C of table 9. The results are weaker, not surprisingly, due to the large loss of misreporting events. Nonetheless, we continue to find evidence that *topic* is incrementally valuable in detecting misreporting. Specifically, we observe in panel A that the joint *topic* and *F-score* model captures the highest rate of AAER violations at the 90th and 95th percentile cutoffs (29.69% and 17.20%, respectively). When benchmarked against the standalone *F-score* model, we note that *topic* improves classification accuracy by 58% at the 90th percentile (accuracy rate of 29.69% versus 18.83%) and by 71% at the 95th (accuracy rate of 17.20% versus 10.04%). Although the *Style* model dominates at the 90th percentile in the AA sample (panel B), the joint *topic* and *Style* model captures the most restatements at the 95th percentile (7.05%), consistent with our results in table 7. The economic value of *topic* is also salient in the restricted 10-K/A sample (panel C). Here the three-vector model performs best in flagging high-risk observations at the 95th and 99th percentile cutoffs. In sum, the results in table 9 corroborate our inferences regarding the economic value of disclosure topics in detecting misreporting.

5.2 ALTERNATIVE CLASSIFICATION OF IRREGULARITY RESTATEMENTS

We use a second approach to identify irregularities from the AA database based on the types of accounting errors corrected by the restatement. This approach mirrors the classification procedure used by Larcker and Zakolyukina [2012] and Zakolyukina [2018]. In these studies, restatements related to revenue recognition and *serious* core expense errors are treated as irregularities, because these corrections are generally associated with negative stock returns when revealed to investors (Palmrose, Richardson, and Scholz [2004], Scholz [2008]). We therefore redefine our *misreport* variable in the AA sample based on whether the restatement corrects a

---

phrases and sentences associated with these topics in Zales' filings for 2004 to 2009 indicate that the discussions pertained largely to advertising and marketing expenses.

[30] We refer back to panel B of figure 1 to illustrate this sample restriction. Assume that the firm depicted in the figure had a third misreporting event in 1999. The firm would be excluded from the out-of-sample prediction for 1999 because it had two misreporting events over the preceding five-year window (1994–1998).

TABLE 9

Out-of-Sample Classification Performance of topic*: Controlling for Repeat Offenders*

Panel A: Classification of AAERs

	50th Percentile		90th Percentile		95th Percentile		NDCG@k
Prediction Model	%	Count	%	Count	%	Count	99th
topic	65.37	92	15.90	21	2.47	7	0.076
F-score	68.41	109	18.83	37	10.04	23	0.141
Style	61.57	94	10.77	17	2.27	6	0.000
topic and *F-score*	67.19	106	29.69	42	17.20	26	0.162
topic and *Style*	68.87	103	11.22	20	3.09	9	0.076
F-score and *Style*	68.61	111	17.93	40	12.78	26	0.162
topic, *F-score*, and *Style*	67.62	108	27.65	44	12.04	24	0.172

Panel B: Classification of AA irregularities

	50th Percentile		90th Percentile		95th Percentile		NDCG@k
Prediction Model	%	Count	%	Count	%	Count	99th
topic	48.03	48	9.45	9	5.49	5	0.000
F-score	53.54	53	7.09	8	2.91	4	0.000
Style	58.07	59	12.02	11	5.49	5	0.048
topic and *F-score*	44.73	45	8.40	9	2.91	4	0.079
topic and *Style*	44.46	46	10.19	10	7.05	7	0.048
F-score and *Style*	48.25	49	9.75	11	4.05	6	0.079
topic, *F-score*, and *Style*	48.81	51	8.40	9	5.48	6	0.079

Panel C: Classification of 10-K/A irregularities

	50th Percentile		90th Percentile		95th Percentile		NDCG@k
Prediction Model	%	Count	%	Count	%	Count	99th
topic	64.02	361	18.13	95	9.57	52	0.087
F-score	57.89	321	15.82	89	9.53	52	0.112
Style	68.63	392	23.83	133	12.38	69	0.063
topic and *F-score*	62.46	348	19.35	108	10.33	58	0.070
topic and *Style*	69.21	392	22.15	127	13.98	77	0.081
F-score and *Style*	68.46	389	24.64	139	14.05	74	0.112
topic, *F-score*, and *Style*	68.37	388	24.47	137	14.37	79	0.130

This table reports the out-of-sample classification accuracy of our prediction models after removing misreporting events by repeat-offender firms from the prediction years. We classify a firm as a repeat offender if the firm has a misreporting event in both the estimation window and the respective prediction year. For such cases, we remove the misreporting event from the prediction year, while retaining the events in the estimation window. Panel A reports the results for the AAER sample, whereas panels B and C present results for the Audit Analytics (AA) and 10-K/A irregularity samples, respectively. In each panel, we report the average annual percentage (%) and total number (Count) of misreported 10-K filings that are accurately classified as misreported by each prediction model at the 50th, 90th, and 95th percentiles of the predicted probability scores. The final column of each panel presents the NDCG@k score for each prediction model, where k is the 99th percentile or the top 1% of the predicted probability scores. The NDCG@k measure evaluates the ranking quality of each prediction model and ranges from 0 to 1, with higher values indicating greater classification performance.

revenue recognition or serious core expense error.[31] The alternative definition identifies 1,751 misreported filings or 5.46% of the AA sample from 2000 to 2012.

The replicated results for the AA sample (not reported) are consistent with our primary findings, though the performance of the prediction models is weaker overall. This decline in predictive ability could be attributable to a higher proportion of immaterial errors being captured by our alternative definition. (Only 59% of this sample is comprised of Big R restatements, compared to 73% in our primary sample.) Nonetheless, *topic* continues to be economically important as classification accuracy for high-risk observations significantly improves when *topic* is added to our benchmark models.

5.3 ROBUSTNESS TESTS

We conduct a series of sensitivity checks for our primary results. Our first set of tests examines the sensitivity of our results to variations in the frequency of misreporting events over our sample period and to the inclusion of technical proxy amendments in our 10-K/A irregularity sample. We then assess the usefulness of *topic* in detecting financial restatements attributable to unintentional misapplications of GAAP (i.e., purely accounting errors). Next, we reestimate *topic* using only narratives from the MD&A section, instead of the full text of the 10-K filings. We also change the regression form to a L1 regularized logit model to alleviate concerns of potential overfitting. Lastly, we replicate our analyses using the raw *topic* proportions of each filing (as opposed to the industry-normalized *topic* proportions) and an expanded set of financial and textual style variables taken from prior studies (Beneish [1997, 1999], Cecchini et al. [2010a, b]). The results from these robustness tests are consistent with our reported evidence. We discuss each test in appendix A.6 of the online appendix.

6. *Conclusion*

We employ a sophisticated machine learning tool to identify and quantify *what* is being disclosed in annual report filings (as opposed to *how* it is being disclosed). We then develop a unique measure, labeled *topic*, that quantifies the thematic content of annual report filings and the attention devoted to each topic. Drawing on the management disclosure and communications literatures, we assess whether *topic* is incrementally informative in predicting intentional misreporting, compared to standard financial and textual style measures.

---

[31] The serious core expense errors include cost of sales errors; liabilities, accounts payable, and accrual estimation failures; and improper capitalization of expenditures. We find similar results if we expand the list to include depreciation or amortization errors; improper recording of payroll, selling, general, and administrative expenses; deferred stock-based and executive compensation errors; and lease and leasehold errors.

Using SEC AAERs and irregularities drawn from financial restatements and annual filing amendments, we find that our *topic* measure provides significant incremental predictive power over commonly used financial statement and textual style measures. Specifically, out-of-sample prediction models that incorporate *topic* outperform models based solely on financial and textual measures. Further, our results reveal that *topic* is incrementally and economically valuable in detecting above-normal and high-risk misreporting events, improving prediction accuracy by as much as 59% in the case of SEC AAERs and 50% for irregularity restatements. Our results are robust to a battery of sensitivity checks, including alternative definitions of *topic*, an alternative identification of irregularity restatements, time variations in misreporting, and additional financial and textual variables.

APPENDIX A

Identification of 10-K/A Irregularities

We conduct an automated text search of amended 10-K filings (10-K/As) to identify material misrepresentations or disclosure omissions that are seemingly intentional. We download and parse all 10-K/A filings from 1994 to 2012 available through the SEC EDGAR FTP site (see appendix A.1 of the online appendix for our parsing methodology). We then use regular expressions to search for specific phrases (in any capitalization) based on the classification criteria set forth in Hennes, Leone, and Miller [2008]. If no corresponding phrase is found, we categorize the filing amendment as stemming from an unintentional reporting problem. The search phrases for each classification criterion are laid out below. The "∗" symbol indicates truncated words, whereas "… " indicates the inclusion of other text.

1. Variants of the words "fraud" or "irregularity": "… fraud∗ …," "… irregular∗ …," "… materially false and misleading …," "… violat∗ of federal securities laws …," "… violat∗ securities exchange act … "
2. Presence of related SEC or Department of Justice (DOJ) investigations: "… sec … investigat∗ …," "… investigat∗ … sec …," "… securities and exchange commission … investigat∗ …," "… investigat∗ … securities and exchange commission …," "… doj … investigat∗ …," "… investigat∗ … doj …," "… department of justice … investigat∗ …," "… investigat∗ … department of justice …," "… attorney general … investigat∗ …," "… investigat∗ … attorney general …," "… u∗s∗ attorney … investigat∗ …," "… investigat∗ … u∗s∗ attorney … "
3. Presence of related independent investigations: "… forensic account∗ …," "… forensic investigat∗ …," "… independent∗ … investigat∗ …," "… investigat∗ … independent …," "… retain∗ … special legal counsel …," "… audit committee … retain∗ …," "… retain∗ … audit committee …," "… audit committee … investigat∗ …," "… investigat∗ … audit committee …," "… for-

mer independent auditors . . . ," " . . . forensic or other account∗ . . . ," " . . . retain∗ . . . independent legal counsel . . . "

APPENDIX B

Variable Definitions

Panel A: Quantitative financial statement and stock return variables

Variable	Definition
log(*Total Assets*)	Log of total assets
RSST Accruals	The sum of changes in working capital accruals, long-term operating assets, and long-term operating liabilities, scaled by total assets; following Richardson et al. [2005]
Δ*Receivables*	Change in accounts receivable scaled by average total assets
Δ*Inventory*	Change in inventory scaled by average total assets
%Soft Assets	Percent of total assets excluding PP&E and cash and cash equivalents
Δ*Cash Sales*	Percentage change in cash sales, where cash sales is measured as total sales minus the change in accounts receivable
Δ*Return On Assets*	Change in income before tax, scaled by average total assets
Actual Issuance	An indicator variable coded as 1 if the firm issued debt or equity securities during the year, 0 otherwise
Operating Leases	An indicator variable coded as 1 if future operating lease obligations are greater than zero, 0 otherwise
Book-To-Market	The ratio of total common equity to the market value of equity, where market value is computed as total common shares outstanding multiplied by closing share price at the fiscal year end
$Lag(Mkt - AdjReturn)$	The previous fiscal year's annual buy-and-hold return inclusive of delisting returns minus the annual buy-and-hold value-weighted market return for the same period
Merger	An indicator variable coded as 1 if the firm completed a merger or acquisition during the current fiscal year, 0 otherwise
Big N Auditor	An indicator variable coded as 1 if the firm was audited by a Big N auditor in the current fiscal year, 0 otherwise.
Mid-size Auditor	An indicator variable coded as 1 if the firm was audited by a mid-size auditor (BDO, Grant Thornton, or McGladrey) during the current fiscal year, 0 otherwise.
TotFinancing	Net cash flow from financing activities, scaled by average total assets
ExanteFinancing	An indicator variable coded as 1 if cash flow from operations minus the prior three year average of capital expenditures, scaled by total current assets is less than −0.5, 0 otherwise
Restructuring	An indicator variable coded as 1 if the firm reported nonzero restructuring charges during the current fiscal year, 0 otherwise

(*Continued*)

APPENDIX B—*Continued*

Panel B: Textual style variables

Variable	Definition
log (*Bullets*)	Log of the number of bullets used in the 10-K filing
Header	The number of characters in the SEC header of the 10-K filing
Newlines	The number of excess newlines in the 10-K filing
Tags	The length of all HTML tags used in the 10-K filing
Parsed Size	The number of characters in the 10-K filing after parsing (see appendix A.1 of the online appendix for the parsing methodology)
Sentence Length	Mean sentence length in words
Word Stddev	Standard deviation of word length
Paragraph Stddev	Standard deviation of paragraph length
Repetitions	The mean number of times each sentence is repeated in the parsed 10-K filing
Sentence Stddev	Standard deviation of sentence length
Type Token Ratio	A measure of vocabulary variation defined as $\frac{UW}{W}$, where UW is the number of unique words in the document and W is the total number of words in the document
Coleman-Liau Index	The Coleman-Liau Index measured as $5.88 \times \frac{C}{W} - 29.6 \times \frac{S}{W} - 15.8$, where C is the total number of characters in the document (excluding spacing and punctuation), W is the total number of words, and S is the total number of sentences
Fog	The Gunning Fog Index measured as $0.4\left(\frac{W}{S} + 100 \times \frac{W'}{W}\right)$, where W' is the number of complex words (three or more syllables) in the document
%Active Voice	The percentage of sentences written in active voice
%Passive Voice	The percentage of sentences written in passive voice
%Negative	The percentage of negative words in the document based on the Loughran and McDonald [2011] dictionary
%Positive	The percentage of positive words in the document based on the Loughran and McDonald [2011] dictionary
All Caps	The number of words in all capital letters with at least two letters
Exclamation Points	The number of exclamation points in the parsed 10-K filing
Question Marks	The number of question marks in the parsed 10-K filing

REFERENCES

Amel-Zadeh, A., and J. Faasse. "The Information Content of 10-K Narratives: Comparing MD&A and Footnotes Disclosures." Working paper, University of Cambridge, 2016.

Anaya, L. H. "Comparing Latent Dirichlet Allocation and Latent Semantic Analysis as Classifiers." Dissertation, University of North Texas, 2011.

Ashbaugh-Skaife, H.; D. W. Collins; W. R. Kinney Jr.; and R. LaFond. "The Effect of SOX Internal Control Deficiencies and Their Remediation on Accrual Quality." *The Accounting Review* 83 (2008): 217–50.

Bao, Y., and A. Datta. "Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures." *Management Science* 60 (2014): 1371–91.

Bao, Y.; B. Ke; B. Li; Y. Julia Yu; and J. Zhang. "Detecting Accounting Frauds in Publicly Traded U.S. Firms Using a Machine Learning Approach." *Journal of Accounting Research* 58 (2020): 199–235.

Bauguess, S. W. "Use of AI and Machine Learning in Market Risk Assessment." Remarks at the Practicing Law Institute 2018 SEC Speaks Conference, February 2018.

BEASLEY, M. S. "An Empirical Analysis of the Relation Between the Board of Director Composition and Financial Statement Fraud." *The Accounting Review* 71 (1996): 443–65.

BEASLEY, M. S.; J. V. CARCELLO; D. R. HERMANSON; AND P. D. LAPIDES. "Fraudulent Financial Reporting: Consideration of Industry Traits and Corporate Governance Mechanisms." *Accounting Horizons* 14 (2000): 441–54.

BELLSTAM, G.; S. BHAGAT; AND J. A. COOKSON. "A Text-Based Analysis of Corporate Innovation." Working paper, University of Colorado, 2019.

BENEISH, M., AND P. VORST. "The Cost of Fraud Prediction Errors." Working paper, Indiana University, 2019.

BENEISH, M. D. "Detecting GAAP Violation: Implications for Assessing Earnings Management Among Firms with Extreme Financial Performance." *Journal of Accounting and Public Policy* 16 (1997): 271–309.

BENEISH, M. D. "The Detection of Earnings Manipulation." *Financial Analysts Journal* 55 (1999): 24–36.

BLEI, D. M. "Probabilistic Topic Models." *Communications of the ACM* 55 (2012): 77–84.

BLEI, D. M.; A. Y. NG; AND M. JORDAN. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (2003): 993–1022.

BLOOMFIELD, R. "Discussion of: Annual Report Readability, Current Earnings, and Earnings Persistence." *Journal of Accounting and Economics* 45 (2008): 248–52.

BONSALL, S. B.; Z. BOZANIC; AND K. J. MERKLEY. "What Do Forward and Backward-Looking Narratives Add to the Informativeness of Earnings Press Releases?" Working paper, Pennsylvania State University, 2014.

BOUKUS, E., AND J. V. ROSENBERG. "The Information Content of FOMC Minutes." Working paper, Federal Reserve Bank of New York, 2006.

BOZANIC, Z.; D. T. ROULSTONE; AND A. VAN BUSKIRK. "Management Earnings Forecasts and Other Forward-Looking Statements." *Journal of Accounting and Economics* 65 (2018): 1–20.

BRAZEL, J. F.; K. L. JONES; AND M. F. ZIMBELMAN. "Using Nonfinancial Measures to Assess Fraud Risk." *Journal of Accounting Research* 47 (2009): 1135–66.

BROWN, S. V., AND J. W. TUCKER. "Large-Sample Evidence on Firms Year-Over-Year MD&A Modifications." *Journal of Accounting Research* 49 (2011): 309–46.

BUSHEE, B. J.; I. D. GOW; AND D. J. TAYLOR. "Linguistic Complexity in Firm Disclosures: Obfuscation or Information?" *Journal of Accounting Research* 56 (2018): 85–121.

CALL, A. C.; G. S. MARTIN; N. Y. SHARP; AND J. H. WILDE. "Whistleblowers and Outcomes of Financial Misrepresentation Enforcement Actions." *Journal of Accounting Research* 56 (2018): 123–71.

CECCHINI, M.; H. AYTUG; G. J. KOEHLER; AND P. PATHAK. "Making Words Work: Using Financial Text as a Predictor of Financial Events." *Decision Support Systems* 50 (2010a): 164–75.

CECCHINI, M.; H. AYTUG; G. J. KOEHLER; AND P. PATHAK. "Detecting Management Fraud in Public Companies." *Management Science* 56 (2010b): 1146–60.

CHANG, J.; S. GERRISH; C. WANG; J. L. BOYD-GRABER; AND D. M. BLEI. "Reading Tea Leaves: How Humans Interpret Topic Models." *Advances in Neural Information Processing Systems* 32 (2009): 288–96.

CHOUDHARY, P.; K. J. MERKLEY; AND K. SCHIPPER. "Do Immaterial Error Corrections Matter?" Working paper, University of Arizona, 2019.

CURME, C.; T. PREIS; H. E. STANLEY; AND H. S. MOAT. "Quantifying the Semantics of Search Behavior Before Stock Market Moves." *Proceedings of the National Academy of Sciences* 111 (2014): 11600–05.

DECHOW, P. M.; W. GE; C. R. LARSON; AND R. G. SLOAN. "Predicting Material Accounting Misstatements." *Contemporary Accounting Research* 28(1) (2011): 17–82.

DECHOW, P. M.; R. G. SLOAN; AND A. P. SWEENEY. "Causes and Consequences of Earnings Manipulation: An Analysis of Firms Subject to Enforcement Actions by the SEC." *Contemporary Accounting Research* 13 (1996): 1–36.

DOUGLAS, K. M., AND R. M. SUTTON. "Effects of Communication Goals and Expectancies on Language Abstraction." *Journal of Personality and Social Psychology* 84 (2003): 682–96.

DOYLE, J.; W. GE; AND S. MCVAY. "Determinants and Weaknesses in Internal Control over Financial Reporting." *Journal of Accounting and Economics* 44 (2007): 193–223.

DYER, T.; M. LANG; AND L. STICE-Lawrence. "The Evolution of 10-K Textual Disclosure: Evidence from Latent Dirichlet Allocation." *Journal of Accounting and Economics* 64 (2017): 221–45.

EAGLESHAM, J. "Accounting Fraud Targeted." *Wall Street Journal*, May 27, 2013.

EICKHOFF, M., AND N. NEUSS. "Topic Modelling Methodology: Its Use in Information Systems and Other Managerial Disciplines". In *25th European Conference on Information Systems (ECIS)* 2017. pp. 1327–47.

FARBER, D. B. "Restoring Trust After Fraud: Does Corporate Governance Matter?" *The Accounting Review* 80 (2005): 539–61.

FEROZ, E. H.; K. J. PARK; AND V. PASTENA. "The Financial and Market Effects of the SEC's Accounting and Auditing Enforcement Releases." *Journal of Accounting Research* 29 (1991): 107–42.

FILES, R. "SEC enforcement: Does Forthright Disclosure and Cooperation Really Matter?" *Journal of Accounting and Economics* 53 (2012): 353–74.

FISHKIN, R. "What SEOs Need to Know About Topic Modeling & Semantic Connectivity." *Moz*, October 2014. Available at https://moz.com/blog/topic-modeling-semantic-connectivity-whiteboard-friday.

GOEL, S., AND J. GANGOLLY. "Beyond the Numbers: Mining the Annual Reports for Hidden Cues Indicative of Financial Statement Fraud." *Intelligent Systems in Accounting, Finance, and Management* 19 (2012): 75–89.

GOEL, S.; J. GANGOLLY; S. R. FAERMAN; AND O. UZUNER. "Can Linguistic Predictors Detect Fraudulent Financial Filings." *Journal of Emerging Technologies in Accounting* 7 (2010): 25–46.

GUAY, W.; D. SAMUELS; AND D. TAYLOR. "Guiding Through the Fog: Financial Statement Complexity and Voluntary Disclosure." *Journal of Accounting and Economics* 62 (2016): 234–69.

HENNES, K. M.; A. J. LEONE; AND B. P. MILLER. "The Importance of Distinguishing Errors from Irregularities in Restatement Research: The Case of Restatements and CEO/CFO Turnover." *The Accounting Review* 83 (2008): 1487–519.

HOBERG, G., AND C. M. LEWIS. "Do Fraudulent Firms Produce Abnormal Disclosure?" *Journal of Corporate Finance* 43 (2017): 58–85.

HOBSON, J. L.; W. J. MAYEW; AND M. VENKATACHALAM. "Analyzing Speech to Detect Financial Misreporting." *Journal of Accounting Research* 50 (2012): 349–92.

HOFFMAN, M.; F. R. BACH; AND D. M. BLEI. "Online Learning for Latent Dirichlet Allocation." *Advances in Neural Information Processing Systems* 32, (2010): 856–64.

HUANG, A.; R. LEHAVY; A. ZANG; AND R. ZHENG. "Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach." *Management Science* 64 (2018): 2833–55.

JANES, H.; G. LONGTON; AND M. PEPE. "Accommodating Covariates in ROC Analysis." *The Stata Journal* 9 (2009): 17–39.

KARPOFF, J. M.; A. KOESTER; D. S. LEE; AND G. S. MARTIN. "Proxies and Databases in Financial Misconduct Research." *The Accounting Review* 92 (2017): 129–63.

LARCKER, D. F., AND A. A. ZAKOLYUKINA. "Detecting Deceptive Discussions in Conference Calls." *Journal of Accounting Research* 50 (2012): 495–540.

LEWIS, C. M. "Keynote Address." In *The 26th XBRL International Conference* (2013, April).

LI, F. "Annual Report Readability, Current Earnings, and Earnings Persistence." *Journal of Accounting and Economics* 45 (2008): 221–47.

LI, F. "The Information Content of Forward-Looking Statements in Corporate Filings: A Naïve Bayesian Machine Learning Approach." *Journal of Accounting Research* 48 (2010a): 1049–102.

LI, F. "Textual Analysis of Corporate Disclosures: A Survey of the Literature." *Journal of Accounting Literature* 29 (2010b): 143–65.

LOUGHRAN, T., AND B. MCDONALD. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *The Journal of Finance* 66 (2011): 35–65.

LOUGHRAN, T., AND B. MCDONALD. "Textual Analysis in Accounting and Finance: A Survey." *Journal of Accounting Research* 54 (2016): 1187–230.

MCLEAN, B. "Is Enron Overpriced?" *Fortune Magazine*, March 2001.

MURPHY, M., AND K. TYSIAC. "Data Analytics Helps Auditors Gain Deep Insight." *Journal of Accountancy* (2015, April/May): 52–58.

NEWMAN, M. L.; J. W. PENNEBAKER; D. S. BERRY; AND J. M. RICHARDS. "Lying Words: Predicting Deception from Linguistic Styles." *Personality and Social Psychology Bulletin* 29 (2003): 665–75.

PALMROSE, Z.; V. J. RICHARDSON; AND S. SCHOLZ. "Determinants of Market Reactions to Restatement Announcements." *Journal of Accounting and Economics* 37 (2004): 59–89.

PEROLS, J. L.; R. M. BOWEN; C. ZIMMERMANN; AND B. SAMBA. "Finding Needles in a Haystack: Using Data Analytics to Improve Fraud Detection." *The Accounting Review* 92 (2017): 221–45.

PURDA, L., AND D. SKILLICORN. "Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection." *Contemporary Accounting Research* 32 (2015): 1193–223.

QUINN, K. M.; B. L. MONROE; M. COLARESI; M. H. CRESPIN; AND D. R. RADEV. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54 (2010): 209–28.

RENNEKAMP, K. "Processing Fluency and Investor's Reactions to Disclosure Readability." *Journal of Accounting Research* 50 (2012): 1319–54.

RICHARDSON, S. A.; R. G. SLOAN; M. T. SOLIMAN; AND I. TUNA. "Accrual Reliability, Earnings Persistence and Stock Prices." *Journal of Accounting and Economics* 39 (2005): 437–85.

ROGERS, J. L.; A. V. BUSKIRK; AND S. L. ZECHMAN. "Disclosure Tone and Shareholder Litigation." *The Accounting Review* 86 (2011): 2155–83.

SCHOLZ, S. *The Changing Nature and Consequences of Public Company Financial Restatements*. Washington, D.C.: U.S. Department of the Treasury, 2008.

ZAKOLYUKINA, A. A. "How Common Are Intentional GAAP Violations? Estimates from a Dynamic Model." *Journal of Accounting Research* 56 (2018): 5–44.