What are you saying? Using topic to detect financial misreporting*

Nerissa C. Brown
Associate Professor
University of Delaware
ncbrown@udel.edu

Richard M. Crowley
Assistant Professor
Singapore Management University
rcrowley@smu.edu.sg

W. Brooke Elliott
Ernst & Young Distinguished Professor
University of Illinois
wbe@illinois.edu

March 2018

*We thank an anonymous reviewer, Phil Berger (the editor), 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 - CUNY, Carnegie Mellon University, Columbia University, HKUST, University of Illinois, U.S. Securities and Exchange Commission (Division of Economic Risk Analysis), Virginia Tech, 2015 AAA FARS Mid-year Meeting, 2015 AAA Annual Meeting, 2015 Conference on Convergence of Financial and Managerial Accounting Research, 2016 Conference on Investor Protection, Corporate Governance, and Fraud Prevention, and the 2016 Conference on Financial Economics and Accounting for their 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, Jill Santore, and Jingpeng Zhu for excellent research assistance.

Brown is serving as a visiting academic fellow in the Office of the Chief Accountant at the U.S. Securities and Exchange Commission. The Securities and Exchange Commission disclaims responsibility for any private publication or statement of any SEC employee or Commissioner. This article expresses the authors' views and does not necessarily reflect those of the Commission, the Commissioners, or members of the staff.

What are you saying? Using topic to detect financial misreporting

Abstract: This study uses a machine learning technique to assess whether the thematic content of financial statement disclosures (labeled as *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 the 1994 to 2012 period. We find that the algorithm produces a valid set of semantically meaningful topics that are predictive of financial misreporting based on samples of SEC enforcement actions (AAERs) and irregularity restatements arising from intentional GAAP violations. Our out-of-sample tests indicate that *topic* significantly improves the detection of financial misreporting when added to models based on commonly-used financial and textual style variables. Furthermore, we find that models including *topic* outperform traditional models when predicting long-duration misstatements. These results are robust to alternative *topic* definitions and regression specifications and various controls for firms with repeated instances of financial misreporting.

Keywords: Topic, 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 in assessing the likelihood of financial misreporting.[1] Detection models of financial misreporting 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. [2018]). One drawback of this approach is that financial misreporting can go undetected for multiple periods since misreporting firms often manipulate performance metrics and accounting transactions to blend in better with either their peers or the firm's own past performance (Lewis [2013]).[2] 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 warning signs of misreporting (e.g., Loughran and McDonald [2011], Hobson, Mayew, and Venkatachalam [2012], Larcker and Zakolyukina [2012], Cecchini et al. [2010a], Purda and Skillicorn [2015]).[3]

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

[1] We use the terms *misreporting*, *misstatement*, *manipulation*, and *irregularity* interchangeably to refer to intentional violations of Generally Accepted Accounting Principles (GAAP). Following Hennes, Leone, and Miller [2008], we refrain from using the term *fraud* since, in a legal sense, accounting misstatements are considered fraudulent only if users rely on the information to their detriment.

[2] In line with this observation, Dechow et al. [2011] find that several financial measures are not significantly different in misreporting years compared to years prior to the manipulation.

[3] Specific features analyzed in prior work include disclosure tone (Loughran and McDonald [2011], Rogers, Buskirk, and Zechman [2011]), vocal cues of cognitive dissonance (Hobson, Mayew, and Venkatachalam [2012]), deceptive words and language cues (Larcker and Zakolyukina [2012]), machine-learned dictionaries of discriminatory words and phrases (Cecchini et al. [2010a], Goel et al. [2010], Purda and Skillicorn [2015]), and measures of readability and textual complexity (Humpherys et al. [2011], Goel and Gangolly [2012]).

narratives. This approach departs from prior text-based research by focusing on *what* is being disclosed in management communications rather than *how* content is disclosed. Using this unique *topic* measure, we evaluate the common types of topics discussed in the annual reports of misreporting firms and how these disclosure topics change over time. More importantly, we investigate the incremental predictive power of *topic* in detecting accounting misstatements out-of-sample relative to a collection of financial statement and textual style measures used in prior work.

Our focus on the thematic content of financial statement filings draws on the communications and management disclosure literature, which 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 (see e.g., Hoberg and Lewis [2017]). This literature also argues that textual features such as tone and word usage can be difficult to classify as deceptive since 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 this sense, the content of the disclosure and the attention devoted to each topic may be better predictors of 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 statement and stock market variables, given that these measures are typically backward-looking and have been shown to be less efficient in predicting misstatements compared to language-based measures (e.g., Cecchini et al. [2010a], Larcker and Zakolyukina [2012], Goel and Gangolly [2012], Purda and Skillicorn [2015]).

To generate our *topic* measure, we employ a Bayesian topic modeling algorithm developed by Blei, Ng, and Jordan [2003], termed Latent Dirichlet Allocation (LDA). Similar to factor or cluster analysis, the LDA algorithm is an unsupervised and unstructured probabilistic

model that "learns" or discovers the latent thematic structure of words within a corpus of documents.[4] 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 the model does not require predetermined word dictionaries or topic categories and instead relies on the basic observation that words frequently appearing together tend to be semantically related. This reduces researcher bias as (preconceived) knowledge of document content does not affect the topic classifications.[5] Furthermore, the algorithm is able to classify the content of large collections of textual narratives—a task that would be infeasible to perform manually on large samples of financial statements.

We derive our topic measures using a comprehensive sample of 131,528 10-K filings issued by U.S. firms over the 1994 to 2012 period. 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 described in Li [2008]. We run the LDA algorithm on the parsed filings using moving five-year windows over our sample period. This time-series approach allows the topic categories to change over time as we expect time-varying factors to influence management communications as well as the ability of thematic content to detect financial misreporting.[6] 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 refer to these topic proportions as *topic*.

We use two approaches to identify filings containing intentional GAAP violations. The first method relies on the Dechow et al. [2011] dataset of SEC Accounting and Auditing

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

[5]While the LDA model 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 probabilities (i.e., word weightings). Thus, some amount of human discretion is unavoidable when applying LDA as a topic classification tool.

[6]Consistent with this notion, Dyer, Lang, and Stice-Lawrence [2017] reports positive time trends in the quantity of 10-K disclosure related to mandatory compliance topics such as fair value accounting, internal controls, and risk factors.

Enforcement Releases (AAERs) issued over our sample period.[7] This data provides specific details on firms subject to SEC enforcement actions for alleged accounting misstatements. We classify 10-K filings as misstated if the AAER sample identifies material GAAP violations that affected the annual reporting period. Our second approach involves an automated search for financial restatements arising from intentional misreporting (hereafter referred to as irregularity restatements).[8] Using the classification criteria detailed in Hennes, Leone, and Miller [2008], we classify the unamended 10-K filing as misstated if our search tool identifies the following in the amended filing: 1) variants of the words "fraud" or "irregularity" in describing the misstatement and 2) references to restatements stemming from investigations by the SEC, Department of Justice (DOJ), or independent parties.

While there is some overlap between our AAER and restatement samples, each data source has its unique advantages and disadvantages. As Dechow et al. [2011] note, the AAER sample provides researchers with high confidence of intentional misreporting since the SEC typically targets firms where there is strong evidence of material misstatements made with the intent of misleading investors. However, one drawback is that many misstatement events are not pursued by the SEC due to resource constraints (Files [2012]). In addition, cases pursued may reflect selection biases arising from the SEC's evaluation process. Irregularity restatements mitigate these limitations since the sample spans a broader set of accounting misstatements. Nonetheless, the restatement sample could introduce other selection issues since our identification method is dependent on how firms disclose or discuss misstatements within the restated filings.

Before conducting our prediction analyses, we evaluate the semantic validity of the LDA output and the ability of the topic categories to detect misstatements in-sample on an annual

[7]We use the updated version of this dataset available from the Center for Financial Reporting and Management at UC Berkeley's Haas School of Business.

[8]We use an automated search to identify irregularity restatements since other data sources such as Audit Analytics and the Government Accountability Office (GAO) database provide less extensive coverage. For instance, restatement data is not available in Audit Analytics for periods prior to 2001, while the GAO database is limited to restatements announced from July 2002 to October 2006.

basis.[9] Using both human and machine-based evaluation procedures, we find that the LDA model produces a coherent set of semantically meaningful topics that captures the underlying economic content of annual report filings. These topics capture content referring to firm performance, risk factors, business transactions, accounting policies, and contingencies, among others. Interestingly, the discussion of particular topics evolve over time, indicating that the content of management communications is quite fluid as argued in Hoberg and Lewis [2017]. The in-sample tests indicate that many of the topic categories are consistently associated with financial misreporting using both the AAER and irregularity restatement samples. For instance, we find that misreporting firms devote more attention to discussing increases in income performance, strategic alliances, and operational growth, while allocating less attention to mergers & acquisitions, hedging activities, legal proceedings, and stock option compensation plans.

Our next set of analyses assess the usefulness of *topic* in detecting misstatements out-of-sample compared to a comprehensive set of quantitative financial statement and market measures as well as a collection of textual style features. Our financial statement and stock market variables stem from an expanded version of the Dechow et al. [2011] *F-score* model and include measures of accrual quality, financial performance, off-balance sheet activities, audit quality, and market-related incentives. Our text-based features include, among others, measures of readability, textual complexity, language voice (active and passive), vocabulary variation, emphasis, and disclosure tone.

Using multiple measures of predictive accuracy, our out-of-sample tests indicate that *topic* provides significant incremental predictive power over commonly-used financial (*F-score*) and textual style (*Style*) metrics. Specifically, models based on *topic* outperform stand-alone models of *F-score* and *Style*, depending on the type of misreporting event. When we evaluate the interplay between all three sets of prediction variables, we find that a joint model

[9]Our in-sample and out-of-sample prediction tests for AAERs (irregularity restatements) are based on a reduced sample of 37,806 (42,314) firm-years over the 1999 to 2010 (1999 to 2012) period. We exclude the 2011 and 2012 years in our AAER tests since no enforcement actions were reported for those two years (see Table 1 for further details).

of *topic* and *F-score* (*Style*) is most predictive of misreporting in the AAER (irregularity restatement) sample. This differential result is not surprising given that the original *F-score* model was built using a sample of AAERs. Additional tests of the classification accuracy of our models indicate that *topic* improves the correct classification of high risk AAER misstatements by roughly 33-46% when added to models of financial and textual style variables. The economic value of *topic* is more modest in the restatement sample with an incremental accuracy rate of 5-6% for high risk misstatements. We also find that *topic* significantly improves the detection of long-duration accounting manipulations that affect multiple annual report filings. Finally, sensitivity checks indicate that our results are robust to alternative regression specifications, *topic* definitions, corrections for potential model overfitting (Perols et al. [2017]), and restricting our topic analysis to the Management Discussion and Analysis (MD&A) section of firms' 10-K filings (Hoberg and Lewis [2017]).

Our study makes several important contributions to the literature. First, we extend prior research on financial misreporting by documenting that the topics discussed in annual financial statement filings are useful in identifying intentional misstatements, either on a stand-alone basis or in combination with standard prediction variables. Second, we expand the burgeoning research in accounting that examines the textual narratives of corporate disclosures. We exploit a robust textual analysis methodology that directly quantifies *what* is being disclosed in annual financial reports (as opposed to *how* information is being disclosed). This content analysis is a significant step forward as it provides a deeper look at the semantic meaning of management disclosures and how disclosure themes are indicative of financial misreporting. Further, our approach takes into consideration the time-varying and fluid nature of management communications, which contrasts with prior work based on word dictionaries that are fairly static and easily identifiable by firms.

Lastly, our study has important practical implications for regulators and accounting professionals, who have begun to implement text-based initiatives aimed at detecting accounting violations and other forms of financial misconduct. For instance, the SEC is developing

computer-powered risk assessment models that leverage text-based tools such as word dictionaries and topic modeling to detect anomalous patterns in registrant filings (Eaglesham [2013]; Bauguess [2018]). Financial statement auditors are also employing textual analytic tools to identify accounting anomalies and assess financial reporting risks (Murphy and Tysiac [2015]). Our results suggest that extracting information on thematic content from disclosure narratives is a rich approach for capturing high-risk accounting activity. Regulators and practitioners should also note that topic analysis is less susceptible to managerial "gaming" relative to other text-based measures, as the words and associated weighting of any topic category are part of an interdependent system. Thus, gaming topic analysis would require a complex and continually evolving unraveling system.[10]

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. Prior research documents that measures of extreme or abnormal financial performance are useful predictors of accounting misstatements. 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, several studies find that accounting misstatements can be explained by stock and debt market pressures (Dechow, Sloan, and Sweeney [1996]) and weaknesses in firms' internal governance or monitoring mechanisms (Beasley [1996], Beasley et al. [2000], Farber [2005]).

Drawing from this body of literature, Dechow et al. [2011] conduct a comprehensive study of AAERs and the detection power of a battery of quantitative financial statement

[10]The natural language field has developed several text simplification tools that can circumvent aspects of language complexity such as readability, language voice, and sentence length (see e.g., Coster and Kauchak [2011]). Many of these tools are now widely available on online platforms (e.g., Rewordify.com, Foxtype.com).

and stock market measures.[11] They find that poor accrual quality, increases in accrual components, declines in returns on assets, high stock returns, and abnormal reductions in the number of employees are strong predictors of accounting misstatements. They also find that misreporting firms conduct aggressive off-balance-sheet and external financing transactions during misstatement periods. Using these variables, Dechow et al. [2011] develop a composite prediction score termed *F-score*. They show that *F-score* is a better predictor of accounting manipulations that are both within and outside of GAAP, relative to traditional models of accrual management.[12]

Despite the usefulness of quantitative metrics in detecting accounting manipulations, prior studies argue that the predictive ability of these measures is quite modest (e.g., 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 tools. The basic premise of these studies is that the linguistic features of management disclosures reveal certain communication patterns that are foretelling of financial misreporting. This stream of research relies on two general approaches to uncover and analyze communication patterns in written disclosures such as financial statement filings and conference call transcripts (see Li [2010b] and Loughran and McDonald [2016] for detailed reviews of these methodologies).

The first approach relies on pre-defined word categorizations (or dictionaries) to investigate the link between accounting misstatements 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

[11]Dechow et al. [2011] do not examine corporate governance and incentive compensation variables because data for these variables are available for only limited samples. Our study follows the same approach to ensure that our results are generalizable to a wide set of firms.

[12]Cecchini et al. [2010b] and Bao et al. [2018] build upon the *F-score* model by using advanced machine learning algorithms and quantitative financial statement ratios to predict accounting misstatements. Both studies find that advanced classification algorithms outperform traditional logistic regressions as used in Dechow et al. [2011]. Nonetheless, in a parallel paper, Cecchini et al. [2010a] find that textual information from 10-K filings significantly improves the detection of accounting misstatements even when the same advanced classification tool is applied. We discuss the results of Cecchini et al. [2010a] in more detail below.

filings are associated with managerial obfuscation as measured by the persistence of earnings. Using the full text of 10-K filings, Loughran and McDonald [2011] find that negative and uncertain language are positively linked to securities lawsuits of alleged accounting improprieties. Rogers, Buskirk, and Zechman [2011] report that firms sued for financial misreporting tend to use substantially more optimistic language in their earnings announcements. Finally, Larcker and Zakolyukina [2012] analyze the transcripts of conference calls and find that deceptive language such as emotion words and anxiety words is a better predictor of misreporting compared to abnormal accrual measures.

The second approach employs machine learning algorithms to discriminate between "bags of words" or textual style markers that are predictive of intentional misreporting. These style markers include textual features such as verbal complexity, readability, document length, and 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 machine learning algorithm termed Support Vector Machine (SVM) to identify or classify accounting misstatements. The SVM approach improves upon prior work as the model learns by example and does not require pre-defined language markers. Cecchini et al. [2010a] uses SVM to generate a dictionary of discriminatory words and phrases that frequently appear in misstated annual reports. Results show that this machine-learned dictionary is more predictive of misstatements, relative to financial statement ratios. Goel et al. [2010] train their SVM tool to recognize misstatements using both word counts and textual features such as disclosure tone, readability, voice (passive versus active), and lexical variety. This alternative SVM approach improves the prediction of accounting manipulations by about 58% compared to baseline bag-of-words models (developed using both SVM and a Naïve Bayes algorithm). Purda and Skillicorn [2015] extend this evidence by analyzing word usage in both annual and quarterly financial reports. Using a sample of AAER firms, they document that a SVM-generated word dictionary outperforms prediction models built using predefined dictionaries as well as models based on accounting and stock market 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 financial reporting. In other words, there is scant evidence on whether the content that managers choose to discuss (or not discuss) is predictive of instances of accounting misstatements. Our study addresses this gap in the literature by using a state-of-the-art topic analysis tool to capture the thematic content of disclosures within firms' financial statements.

2.2 LDA Topic Modeling

We employ a topic modeling approach developed by Blei, Ng, and Jordan [2003], termed Latent Dirichlet Allocation (LDA), to capture the thematic content (i.e., topics) 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 application to various text collections). We use this approach to construct a firm-specific measure of the topics discussed in annual financial statements in a given reporting year. This unique measure (defined as the normalized percent 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.

Topic modeling is relatively new to accounting and finance (see Loughran and McDonald [2016] and Eickhoff and Neuss [2017] for reviews of relevant studies), and our measurement approach is consistent with recent studies that use the LDA technique to investigate various financial reporting and capital market issues. Specifically, Curme et al. [2014] use LDA to identify the semantic topics within the large online text corpus of Wikipedia. The identified topics are then used to determine the link between stock market movements and how frequently Internet users search for the most representative words of each identified topic. Huang et al. [2017] employ LDA topic modeling to compare the thematic content of analyst

reports and the text narrative of conference call transcripts. Consistent with the information discovery role of analysts, Huang et al. find that analyst reports issued immediately after conference calls contain exclusive topics that were not discussed during the calls. Bao and Datta [2014] discover and quantify the various topics discussed in textual risk disclosures from the Item 1A section of annual 10-K filings. The results indicate that about two-thirds of the identified risk topics are uncorrelated with measures of investors' risk perceptions, consistent with the notion that risk disclosures are largely boilerplate. However, of the remaining topics, Bao and Datta [2014] find that discussions of systematic macroeconomic and liquidity risks have an increasing effect on investors' risk perceptions, whereas discussions of diversifiable risks (e.g., operational and information security risks) reduce risk perceptions. Lastly, Dyer, Lang, and Stice-Lawrence [2017] employ LDA analysis to determine the specific topics that contribute to increases in the length of 10-K filings over time. Their results suggest that new FASB standards and SEC mandates regarding fair value, internal controls, and risk factor disclosures account for much of the upward trend in 10-K length.

In a concurrent study, Hoberg and Lewis [2017] use LDA topic modeling and cosine similarity to provide evidence of the content disclosed by firms involved in SEC enforcement actions (AAERs) and to examine underlying incentives for misreporting firms to produce abnormal disclosure content. Focusing on the MD&A section of 10-K filings, Hoberg and Lewis [2017] find that, relative to industry peers, AAER firms disclose abnormal textual content that is common among misreporting firms. Their topic analysis indicates that during misreporting years AAER firms are more likely to grandstand manipulated revenue and R&D expense performance, while providing fewer quantitative details explaining how their performance arises. This evidence is consistent with incentives for misreporting firms to tout strong performance and growth options while concealing broad performance details. In addition, AAER firms tend to under-disclose topics related to liquidity challenges in an attempt to lower cost of capital effects. Lastly, managers tend to disassociate themselves from the firm during misreporting years by disclosing abnormally low levels of content referencing

the manager's overall participation in the firm's vision and strategy.

Our study differs from Hoberg and Lewis [2017] in several respects. First, our study goes beyond disclosure incentives to demonstrate the incremental predictive power of thematic content in detecting accounting misstatements within the broader population of financial statement filers. As a results, our study seeks to provide new insights into the potential benefits of statistical topic analysis in assessing the likelihood of financial 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 empirical analyses. As highlighted in prior studies (e.g., 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 to current text-based methodologies. We therefore extend Hoberg and Lewis [2017] by accounting for the time-varying nature of management disclosure in annual report filings. Specifically, our methodology employs a rolling-window estimation procedure to simultaneously discover and quantify the topics discussed by management in a given reporting year. We use the same rolling-window setup to predict financial misreporting using the topics identified over the five years prior to the manipulation period.

Lastly, the analysis in Hoberg and Lewis [2017] is confined to the text contained in the MD&A section (Item 7), whereas our study considers the thematic content of the entire 10-K filing. While the MD&A section provides a useful setting for examining disclosure content, it does not capture relevant content discussed in other sections of the annual report (Li [2010b]), e.g., risk factors (Item 1A), legal proceedings (Item 3), and executive compensation (Item 11). Moreover, as noted in Loughran and McDonald [2016], an additional drawback of focusing on only one section of the annual report is that companies can strategically shift or (de-)emphasize content across multiple sections.[13] As we will show, topics identified in

[13]In line with this notion, Amel-Zadeh and Faasse [2016] find that management's tone in footnote disclosures is significantly more negative than the tone of the MD&A, especially when firm performance is poor. This result potentially reflects managements' attempt to downplay negative accounting information by putting a more positive spin on the content discussed in the MD&A section.

the MD&A section have lower detection power compared to topics identified from the full annual report.

2.3 Research Questions

Our primary research questions explore the usefulness of thematic content in identifying intentional financial misreporting. We are particularly interested in assessing the *incremental* predictive value of thematic content, relative to the predictive value of traditional quantitative financial measures and textual style characteristics.

Disclosure topics may provide incremental detection ability since the thematic content of financial statements is likely to capture an aspect of managerial deception that is distinct from that of financial metrics and textual style features. Regulatory oversight is more difficult for financial statement narratives, especially at the topic level, thus leaving more room for managers to use disclosure content as a means of obfuscating accounting manipulations or providing misleading information. While prior research has identified a set of deceptive words and cues (see e.g., Newman et al. [2003], Larcker and Zakolyukina [2012]), it is more difficult for regulators to naïvely identify and monitor a set of deceptive topics as these same topics may be benign or informative about true financial performance in other settings or at other points in time. Furthermore, managers engaging in accounting manipulations have discretion to vary not only the specific content discussed throughout the financial statements but also the amount of attention devoted to any given topic. Thus, as argued in Hoberg and Lewis [2017], the flexible nature of disclosure content allows for a broader set of dimensions along which annual report narratives can be used to identify accounting manipulations, relative to quantitative metrics and aggregate measures of textual style features.

In contrast to quantitative financial information, which is primarily backward-looking, prior research finds that a sizable proportion of the textual narratives in earnings announcements and financial statement filings contain forward-looking statements that cover a wide range of topics (see e.g., Li [2010a], Bozanic, Roulstone, and Van Buskirk [2017]). Prior ev-

idence also suggests that management's discussion of forward-looking information responds differently to firm conditions such as poor earnings performance, financial distress, and industry competition, compared to discussions of backward-looking information (Bonsall IV, Bozanic, and Merkley [2014]). Together, these results provide further support to the notion that disclosure topics respond differently to managerial deception and obfuscation, relative to quantitative financial measures.

Our research questions pay particular attention to whether disclosure topics provide informational value beyond textual style characteristics in identifying accounting manipulations. This query is warranted as linguistics research suggests that it is difficult to discern deception or obfuscation from textual features such as tone and abstract language since communications can be flavored by individuals' expectations and motivations even when the intent is to communicate objectively and truthfully (see Douglas and Sutton [2003]). Furthermore, prior accounting research debates whether textual style characteristics such as length, readability, and word usage reflect managerial deception or obfuscation, or simply the inherent complexity of discussing unusual performance or business events (see e.g., Li [2008] and related discussions in Bloomfield [2008] and Loughran and McDonald [2016]). Indeed, recent studies by Guay, Samuels, and Taylor [2016] and Bushee, Gow, and Taylor [2018] suggest that aggregate measures of textual or document complexity can be ambiguous indicators of managerial obfuscation in written and scripted business narratives such as annual reports and conference call presentations. We however note that similar arguments can be made for the topics communicated by management since certain discussions within the annual report can reflect unusual business events and/or new regulatory requirements as opposed to managerial obfuscation (see Dyer, Lang, and Stice-Lawrence [2017]). Nonetheless, in contrast to summary textual measures, the detailed nature of topic analysis allows for greater insight into how the content of annual report narratives and the attention dedicated to each topic vary with the incidence of financial misreporting.

Overall, the above arguments suggest that the topics discussed in annual reports may pro-

vide incremental predictive value beyond quantitative financial metrics and aggregate textual style measures in detecting financial misreporting. However, any incremental predictability is largely an empirical question given the variety of ways that accounting manipulation and deception can manifest in written managerial communications. In addition, financial statement filings are joint outputs of not only management but also legal counsel, leading to disclosure content that is fairly sanitized and boilerplate in some areas (Brown and Tucker [2011], Hoberg and Lewis [2017]). This conservative aspect of the financial reporting process further implies the empirical nature of our arguments. We therefore present our predictions in the form of research questions as follows:

> **Research Question 1**: *Topic provides predictive information beyond that of quantitative financial statement and stock market measures when detecting intentional financial misreporting.*
>
> **Research Question 2**: *Topic provides predictive information beyond that of summary measures of textual features when detecting intentional financial misreporting.*

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. Despite the conservative nature of annual reports, we focus our analysis on these filings because they 1) allow us to maximize the number of firm-year observations in our sample, 2) are comprehensive in their coverage of the firm and its activities throughout the fiscal year, and 3) avoid self-selection biases given their mandatory disclosure status.[14] We download the full text of all 10-Ks available through the SEC EDGAR FTP site from January 1, 1994

[14]We acknowledge that 10-K filings are not always timely sources for detecting financial misreporting. For instance, the filing of the 10-K can lag any occurrence of misstatement by up to a year. Purda and Skillicorn [2015] highlight the added value of including quarterly report narratives in language-based analyses of financial misreporting. However, we choose to exclude quarterly reports from our analyses to ensure consistency in the disclosure content of financial statement filings across firms' reporting periods.

(the first year such data is available) until December 31, 2012 (the final year of the AAER dataset as discussed below). This download yields 131,528 10-K filings by U.S. firms over the 1994 to 2012 period. We use the full set of filings to generate the topic measures as this improves the algorithm's convergence.

We follow Li [2008] in parsing the 10-K filings, but expand this methodology to remove all items included in the documents other than raw text.[15] To remove typos and uncommon terminology, we also restrict the words in the files to match those contained in the standard Unix words dictionary.[16] We describe our full parsing methodology in detail in Appendix A.1 of the online appendix. We gather data on accounting misstatements from the SEC AAER dataset compiled by Dechow et al. [2011] and from disclosures of restatements due to intentional misreporting in amended 10-K filings. We also gather financial statement and stock market data from Compustat and CRSP, respectively. To conduct our prediction analyses, we take the full set of 10-K filings and exclude financial firms (SIC codes 6000-6799) and firm-years with missing Compustat and CRSP data necessary to compute our quantitative financial metrics. We also exclude from our rolling-window prediction tests those firm-years with zero instances of accounting misstatements (for e.g., there are no AAER events in our prediction sample for the 2011 and 2012 years). Table 1 reports the distribution of the observations used in our prediction tests for AAERs and irregularity restatements, respectively. We discuss the construction of these samples in more detail below.

3.1.1 Identifying Intentional Financial Misreporting

We use two data sources to identify instances of intentional financial misreporting. Following Dechow et al. [2011], our first data source relies on SEC AAERs to identify firms engaging in material accounting misstatements. We create an indicator variable (*misreport*) that equals

[15] We construct measures for all textual items removed from the documents, some of which are included in our analyses.

[16] The standard dictionary, provided by the wamerican package in the official Debian repositories, contains 99,171 words. We also conduct robustness checks using no dictionary, the wamerican-huge dictionary, and the wamerican-insane dictionary. These checks confirm that the standard dictionary provides the best model performance in-sample, along with the most coherent topics.

1 for each fiscal year affected by the accounting violation as identified by the SEC, and zero otherwise. We then use this indicator variable to classify those 10-K filings containing alleged GAAP violations. Our second data source is a customized automated search for occurrences of restatements of annual financial statements that are seemingly due to intentional misapplications of GAAP (referred to as "irregularity restatements"). We use the classification scheme discussed in Hennes, Leone, and Miller [2008] to develop a customized identification tool. A manual inspection of a random sample of irregularity restatements indicates that our customized tool performs well in capturing instances of financial misreporting.[17]

To identify irregularity restatements, we download all amended 10-K filings (10-K/As) from the SEC EDGAR FTP site. We gather firm-identifying information for matching purposes from the header (or alternately from the body of the text when the header is missing or incomplete), and then parse the 10-K/A in a manner similar to our parsing of unamended 10-Ks. After parsing the filings, we search the text for direct statements of the occurrence of financial reporting irregularities or narratives referring to the investigation of misstatements by either regulatory or independent parties. Appendix A describes our full search terms. Specifically, we search for phrases such as "fraud," "materially false and misleading," and "violation of federal securities laws" to identify restated annual filings with direct discussion of irregularities. We identify restatements with related regulatory investigations based on narratives referring to investigation by the SEC, the DOJ, or by an Attorney General. Restatements with independent party investigations are classified based on discussions referring to investigations by forensic accountants, the audit committee, or an independent committee, as well as statements referring to the retention of legal counsel over the misstatement. Based on this identification strategy, we classify each 10-K filing as misstated if our search of the corresponding 10-K/A detects narratives reflecting an irregularity as detailed above. We then code *misreport* as 1 for those fiscal years with misstated annual reports; *misreport* equals 0 if there is no amended 10-K for the respective fiscal year or if the amended 10-K

[17] We hand-check a random sample of irregularity restatements identified by the search tool and find that the misstated financial reports contained material and intentional misapplications of GAAP.

filing does not involve an irregularity.

Table 1 reports the frequency of *misreport* by year for both AAERs and irregularity restatements over the 1994 - 2012 period, along with the percent of firm observations with detected misstatements. Consistent with prior research (e.g., Larcker and Zakolyukina [2012], Perols et al. [2017], Bao et al. [2018]), the frequency of detected misreporting is very low with overall rates of 1.21 and 1.65 percent, respectively, for the AAER and irregularity restatement samples.[18] We also observe greater consistency in the rate of restatement events in the post Sarbanes-Oxley era, relative to the rate of AAER events. This evidence is not surprising given the long time lag in the SEC's investigation and processing of enforcement cases (see similar finding in Bao et al. [2018]).

3.2 Empirical Measures

3.2.1 Financial Measures

We draw our quantitative financial statement and market-related variables from the Dechow et al. [2011] *F-score* model (see Model 3 in Table 9 of their paper).[19] These variables capture accrual quality, firm performance, off-balance-sheet activities, and market-based incentives. We augment the *F-score* model with variables capturing firm size, audit quality, and firm involvement in complex business transactions, namely, mergers and acquisitions (M&As) and restructurings. Panel A of Appendix B defines each of the variables outlined below.

The accrual quality measures include an extended definition of working capital accruals (termed RSST accruals) which captures the change in noncash net operating assets (Richardson et al. [2005]).[20] We also measure the change in receivables and the change in inventory

[18]The low rate of misreporting events creates a data rarity issue that can lead to overfitting in prediction modeling (Perols et al. [2017]). We correct for this issue in our robustness tests and find that our results are qualitatively similar (see Section 5.4 for further details).

[19]Our results are robust to the inclusion of standard financial ratios and bankruptcy prediction measures, consistent with prior studies (e.g., Beneish [1997], Cecchini et al. [2010a,b]).

[20]We do not include other measures of discretionary accruals (e.g., modified Jones and performance-matched discretionary accruals) as Dechow et al. [2011] find that these measures perform poorly in detecting accounting manipulation compared to unadjusted accrual measures.

since misstatement of these two accrual components affects widely-used profitability metrics. The percent of soft assets on the balance sheet captures accounting flexibility, and in turn, the room for managerial discretion in changing the measurement assumptions of net operating assets in order to meet short-term performance goals. Our performance measures capture (poor) firm performance as a common determinant of accounting manipulations. These measures include the change in cash sales and the change in return on assets. To gauge the extent to which firms engage in off-balance-sheet financing, we include an indicator variable to identify firm-years with nonzero future operating lease obligations. We use the firm's stock price performance and external financing needs to proxy for market-related pressures to engage in accounting manipulations. These proxies include the book-to-market ratio, market-adjusted return over the prior fiscal year, firm leverage, actual issuance of debt and equity securities in a given firm-year, the net amount of new capital raised, and the ratio of estimated free cash flows to the actual balance of current assets.

Our final set of variables focuses on characteristics that are correlated with firm size and the quality of external and internal monitoring mechanisms. Prior research documents that the quality of audit-firm monitoring is an important predictor of misstatements (e.g., Farber [2005]). We measure audit firm quality using separate indicator variables for whether the firm is audited by a Big N or mid-size auditor in the current fiscal year. Studies also show that internal control weaknesses and misstatement risk are generally higher for firms involved in M&As and restructurings (e.g., Doyle, Ge, and McVay [2007], Ashbaugh-Skaife et al. [2008]). We therefore include indicator variables for M&A and restructuring activities in the current fiscal year.[21] Lastly, larger firms are more likely to invest in monitoring mechanisms that mitigate the occurrence of aggressive accounting activity. We use the log value of total assets to capture potential size effects.

[21]In robustness checks, we replace these measures with indicators for M&A and restructuring activities in the current fiscal year or previous two years. This alternative approach yields qualitatively similar results.

3.2.2 Textual Style Measures

We benchmark our *topic* measure against a comprehensive set of textual features (denoted *Style*) used in prior literature, as well as four additional measures developed from our parsing process. Panel B of Appendix B presents a full list of our *Style* variables and their measurement.[22] Following Li [2008] and Goel et al. [2010], our first group of textual style variables are surface features that proxy for disclosure readability and textual complexity. These variables include the mean and standard deviation of the length of words, sentences, and paragraphs in the 10-K filing, as well as measures of sentence repetition and type-token ratio (see Goel et al. [2010], Li [2014]). We also compute the percent of short and long sentences ($\leq$ 30 or $\geq$ 60 words, respectively), along with two complementary measures of readability: the Gunning Fog Index and the Coleman-Liau Index. These indices are widely used in prior studies to capture disclosure inefficiencies and potential managerial obfuscation (see e.g., Li [2008]).

Our next set of measures comprises deeper linguistic markers such as voice, tone, lexical variety, and disclosure emphasis (see e.g., Goel et al. [2010], Purda and Skillicorn [2015]). We measure language voice as the percentage of sentences with active and passive verbs. We define negative and positive disclosure tone using the word dictionaries constructed in Loughran and McDonald [2011]. We use the type-token ratio (number of unique words scaled by the number of total words) to measure lexical variety. Consistent with Rennekamp [2012], this measure captures the use of superfluous or meaningless words, as a higher ratio indicates a broader vocabulary. We also measure disclosure emphasis based on the number of capitalized words, exclamation points, and question marks within the 10-K filing (Goel and Gangolly [2012]).

[22]Our results are robust to a large vector of alternative textual features. This vector includes 1) a battery of processing measures (a variable for each part removed from the filing), 2) median word, sentence and paragraph lengths (in addition to the already included mean lengths), 3) Harvard IV dictionary measures, 4) six alternative readability measures, 5) a variable capturing every part of speech coded in the Brown corpus, 6) total and tagged word counts, 7) two alternative measures of sentence repetition, and 8) the deviation from the Benford distribution. We find that the majority of these variables are highly correlated with the textual features selected for our primary analyses.

Finally, we construct additional measures derived from our processing of the 10-K filings. These measures include 1) the log of the number of bullets, 2) the length of the SEC mandated header, 3) the number of excess newlines (vertical whitespace) within the filing, and 4) the character length of HTML tags. These measures are exploratory in nature and attempt to control for unobservable factors that may be correlated with our variables of interest as well as other textual measures derived from the 10-K filing.

3.2.3 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].[23] We choose this approach due to its intuitive characteristics and strong performance. LDA is a Bayesian probabilistic model and offers significant theoretical improvements over older data-driven and principle-component-based tools such as Latent Semantic Analysis (LSA). Furthermore, the topic modeling accuracy of LDA is quite strong when compared to human classification of topics or other unsupervised algorithms such as LSA-IDF or LSA-TF.[24] In an experiment, Anaya [2011] finds that humans classify main topics with 94% accuracy, while LDA achieves 84% accuracy. Comparable accuracy statistics for LSA-IDF and LSA-TF are 84% and 59%, respectively. 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, a structured literature review by Eickhoff and Neuss [2017] documents that LDA and other topic models have been successfully applied to business documents in several disciplines including accounting, finance, information systems,

[23]For predictive purposes, McAuliffe and Blei [2008] develop a supervised LDA model (sLDA) which allows each text document to be paired with a response variable that classifies each document. The goal of the sLDA model is to infer disclosure topics that are predictive of the response. The response in our setting would be instances of accounting misstatements. We choose not to use the sLDA model for two reasons. First, the unsupervised LDA model allows us to provide a baseline for the common disclosure topics contained in annual reports, irrespective of misreporting. Second, McAuliffe and Blei [2008] find that the prediction performance of sLDA is equivalent to LDA in text corpora with difficult-to-predict responses. A similar result is likely to hold for misstatements given the rarity of these events.

[24]LSA-IDF and LSA-TF are LSA based measures using a term-document matrix that has undergone a transform: inverse document frequency (IDF) or term frequency (TF), respectively.

and management.[25] Many of these studies use qualitative or quantitative methods (or both) to evaluate the effectiveness of LDA in identifying business-related topics. Qualitatively, Bellstam, Bhagat, and Cookson [2017] find that LDA effectively generates topics from analyst reports that conform well to concepts used to describe innovation activities. Quantitative analyses confirm that these topics are strongly correlated with existing innovation measures and exhibit patterns that fluctuate as expected with temporal and industry variations in innovation. Huang et al. [2017] use a single human coder to classify the topics discussed in the conference call transcripts and analyst reports for three companies in the same industry. They find that the LDA topic assignments are consistent with the topics identified by the human coder in 60-69% of the sentences contained in the transcripts and analyst reports. Using qualitative techniques such as topic labeling and temporal variation in economic events, Huang et al. [2017] also find that LDA reliably captures the underlying economic content of analyst reports and conference calls. Lastly, Dyer, Lang, and Stice-Lawrence [2017] employ the human validation technique recommended in Chang et al. [2009]. Their results show that the LDA topics identified from a large corpus of 10-K narratives are coherent and meaningful when tested against human intuition. Taken together, the results from these studies indicate that the LDA algorithm performs well when classifying the thematic content of financial statement and other business-related narratives.

The LDA model is based on a few simple assumptions. The model assumes a collection of K topics in a given text document and that the vocabulary of each topic is distributed following a Dirichlet distribution, $\beta_K \sim \text{Dirichlet}(\eta)$.[26] 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

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

[26] A Dirichlet distribution is essentially a multivariate generalization of a beta distribution.

associated with multiple topics. This is a distinguishing feature of LDA, as words can have multiple meanings, especially in different contexts. In sum, the LDA approach can be viewed as a probabilistic process that condenses the vocabulary in a collection of documents into a set of topic weights and a dictionary of topics.

We implement the LDA algorithm using a dynamic time-series process since, consistent with prior work (e.g., Dyer, Lang, and Stice-Lawrence [2017]), we expect annual report content to change across time due to factors such as macroeconomic conditions, technological changes in business operations, changes in accounting standards and disclosure requirements (e.g., the 2002 Sarbanes-Oxley Act), and changes in firm management. Consequently, this approach allows us to assess the changing nature of disclosure content and its ability to predict accounting misstatements. Our time-series procedure identifies the topics discussed in each rolling five-year window over our sample period (1994 – 2012). That is, we run the algorithm for the periods 1994 – 1998, 1995 – 1999, 1996 – 2000, and so on. The topics discovered in each window are then used to determine the disclosure content of annual reports issued in the year immediately following each five-year window. This results in a test period of 1999 – 2012 for our prediction analyses. Note that while 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 in our prediction tests.

We follow Hoffman, Bach, and Blei [2010] and implement the algorithm using an "online" or batch variant of LDA. This approach is computationally efficient as it allows us to run the algorithm in small batches (100 filings in our case). We draw the filings in each batch in random order to mitigate overweighting of early years in the online LDA tool. Consistent with Hoffman, Bach, and Blei [2010], we use symmetric Dirichlet distributional parameters of $\alpha = \eta = \frac{1}{20}$ and the learning parameters of $\kappa = \frac{7}{10}$ and $\tau_0 = 1024$. The learning parameter κ controls how quickly old information is forgotten, while parameter τ_0 downweights early iterations of the model. Hoffman, Bach, and Blei [2010] document that these distributional

and learning parameter settings are optimal when categorizing articles from the science journal *Nature*, as well as categorizing text from Wikipedia. We then set the algorithm to identify 31 topics in each five-year window. We select 31 topics since simulated results indicate that this number of topics is optimal in capturing the occurrence of irregularity restatements (see Appendix A.2 of the online appendix for a description of this simulation).[27]

Next, we pre-process the parsed 10-K filings by first removing stop words. Stop words are those deemed irrelevant for our text-based measures because they occur either frequently (e.g., 'the', 'an', 'is') or are too infrequent to be of predictive value (such cases were often garbled text or misspellings in the 10-K filings). Because our analysis uses rolling five-year windows, we generate our stopwords on matching five-year windows to avoid potential look-ahead biases. We remove three types of stopwords: 1) the most frequent words appearing in each rolling five-year window until we have removed 60% of all word occurrences in the window, 2) words that occur less than 1,100 times in the window, and 3) words that occur in less than 100 filings. These parameters are also derived in our simulation (see Appendix A.2 of the online appendix).

We run the LDA algorithm on the pre-processed filings, generating 31 topics in each rolling 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 following the five-year window (i.e., the word weights for topics identified in the 1994 – 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 for the filing. We then scale the topic weights by the sum of the weights of all topics identified in the filing. This procedure generates the proportion of the content of each document that is associated with each topic. We denote these topic proportions as *topic*.

[27] We run the simulation on our irregularity restatement sample given the lower frequency of SEC AAERs.

4 Empirical Results

4.1 Evaluation of LDA Topic Measure

Before investigating our research questions, we assess the validity of the topics inferred by the LDA algorithm. Since LDA is an unsupervised method, it is necessary to evaluate the algorithm's effectiveness in capturing human intuitions about the thematic content of 10-K narratives. We therefore follow prior studies and assess the semantic meaning and interpretability of the inferred topics using both human and machine-based evaluation methods (see e.g., Chang et al. [2009], Quinn et al. [2010], Bao and Datta [2014], Huang et al. [2017], Bellstam, Bhagat, and Cookson [2017], Dyer, Lang, and Stice-Lawrence [2017]).[28]

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 the topics provide meaningful economic content. This form of evaluation is qualitative and follows from several studies including Quinn et al. [2010], Bao and Datta [2014], Bellstam, Bhagat, and Cookson [2017], Hoberg and Lewis [2017], and Huang et al. [2017]. However, one limitation of this approach is that it naturally involves human discretion given the manual process of topic labeling as described below. Nonetheless,

[28] An alternative validation approach is the use of human coders to identify the topics discussed in a subset of our 10-K narratives. The topics identified by the human coders would then serve as a benchmark for assessing the accuracy of the LDA topic assignments. We however refrain from using this approach given the challenging nature of manual topic coding (see Quinn et al. [2010] for a discussion of these challenges). To make human classification tasks feasible, prior studies typically present the human coders with a fixed list of broad potential topics 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 of text. This directed form of human coding is used in both Anaya [2011] and Huang et al. [2017] as a benchmarking method (see Section 3.2.3 for earlier discussion of these studies). Specifically, Huang et al. [2017] present a single human coder with a list of the topic labels inferred from the LDA model prior to commencing the task. Likewise, Anaya [2011] presents coders with a list of broad topic labels from which they selected the topic that best matched the narrative. While ex ante knowledge of potential topic labels 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 topic classification as a validation method. We therefore focus our analyses on evaluating whether our LDA topics have human-identifiable semantic coherence. This approach is superior in that it does not require us to devise a list of benchmark topics or provide our inferred LDA topic labels to human judges beforehand. Our approach is also more feasible and less costly given the time consuming nature of manual topic coding.

we must note that these labels do not influence our empirical analyses, since 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 rolling-window 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. These aggregate topics are referred to 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 several topics do not appear in all 14 windows. 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 topics are achieved when the correlation threshold is set at 11%, resulting in 64 combined topics across our sample period.[29]

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 top 20 most frequent bigrams (i.e., two-word phrases excluding stopwords, numbers, and symbols) are then extracted from the 334 mid-length sentences. These sentences are also sorted based on the cosine similarity between a given sentence and the remaining 333 sentences.

[29]The first pass of this test determined that the optimal correlation threshold ranged between 8% and 18%. We then conduct tests of this threshold range in 0.05% increments to locate the 11% cut-off point. We also compare the combined topics generated by groupings based on Spearman correlation and Euclidean distance. Both of these alternative methods performed poorly due to overweighting on words with low topic weightings, leading to incoherent topic groupings.

We manually review the top 20 bigrams and top 100 mid-length sentences based on cosine similarity, and assign descriptive labels to each of the combined topics.

Appendix C presents a list of the 64 combined topics with 10 representative bigrams per topic and our inferred topic labels.[30] For ease of exposition, we list all industry-specific topics at the end of Appendix C (topics 50 to 64). The reported bigrams exclude redundant phrases (e.g., "millions in," "company also," "in year") and those with similar inferences (e.g., "compared to" and "compared with" in topic 2, or "derivative financial" and "financial derivative" in topic 9). We note that the LDA algorithm performs well in identifying narrative content that is distinctively related to changes in firms' financial performance. For instance, topics 1 and 2 both refer to the firm's income performance compared to prior periods. Examples of top mid-length sentences from topic 1 include the following: "Other income decreased to $11,745,000 in 1999 as compared to $11,882,000 in 1998 and $10,521,000 in 1997" and "Management fee income decreased to $0 as compared to $1.4 million in 1997." Other performance-related topics include segment performance (topics 12 and 40), franchise revenues (topic 17), and general references to quantitative financial statement information (topics 6, 24, 47, and 48). LDA also identifies topics related to complex business transactions and arrangements such as fuel and natural gas purchase commitments (4), derivatives and hedging activities (8 and 29), merger activities (22), R&D partnerships (23), joint ventures (27), strategic alliances (32), and investments in securitized securities (41).

Several topics refer to specific financial statement items and/or their underlying measurement assumptions. These include post-retirement health care cost assumptions (3), account receivables and doubtful accounts (10), long term assets (16), advertising expenses (25), and the measurement of natural gas properties (60). Consistent with Huang et al. [2017], we are able to identify industry-specific topics such as aircraft leasing arrangements (54) and manufacturing operations (54) in the airline industry, measurement of natural gas properties in

[30]The inferred labels for a few topics are overlapping due to only minor differences in the content inferred from the bigrams and mid-length sentences. We treat these topics separately in our empirical analyses to mitigate any noise introduced by our topic aggregation process.

the oil and gas industry (61), as well as general discussions of business risks and operational factors in the agricultural, gaming, mining, marine transportation, and hotel industries. Lastly, as demonstrated in Bao and Datta [2014], LDA effectively discovers content related to common risk factors and contingencies such as foreign currency risks (43), country risks (13 and 26), environmental liabilities and risks (5 and 42), patent infringement and rights (34), and legal proceedings (31).[31]

4.1.2 Word Intrusion Tasks

Our next evaluation method uses "word intrusion" tasks to assess the interpretability or 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 try to identify an unrelated or "intruder" word inserted into a list of words that the LDA model selected as belonging to the same topic. If the set of words from the LDA model is coherent, then the human subjects should easily correctly the intruder word at a rate that is significantly higher than random chance. In other words, a higher identification rate indicates higher interpretability or coherence of the LDA output.

We conduct our word intrusion task using both human-subject as well as machine-based procedures. Given the large number of topics across our rolling 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 first discuss the results of our machine-based tests before turning to our human-subject results.

[31] As a robustness check, we follow Bao and Datta [2014] and assess whether our LDA topics are significantly correlated with investors' risk perceptions as proxied by future stock return volatility. Similar to Bao and Datta [2014], we find that 10-K discussions of macro-level risk factors such as foreign country and currency risks, environmental risks, and commodity risks are positively associated with future return volatility. We also find that discussions of hedging activities, securitizations, and strategic partnerships and joint ventures influence investors' risk perceptions.

We use a word embedding machine algorithm to conduct a series of word intrusion tasks across our topics. Word embedding algorithms are trained to reconstruct linguistic contexts of words. The algorithm creates a vector mapping of the semantics of large text corpuses such that words sharing similar contexts are located in close proximity to each other in the vector space, and pairs of words with analogous relationships are located with similar distances between the words.[32] Once the vector mapping is created, the algorithm will be able to detect unrelated or intruder words similar to a human subject.

To conduct the word intrusion task, we first train the algorithm on the text narratives from the 131,528 10-K filings issued over our sample period, 1994-2012. For robustness, we also train the algorithm on a large corpus of randomly sampled daily editions of the *Wall Street Journal* (*WSJ*) over the sample period from ProQuest. The *WSJ* corpus provides a useful benchmark for the performance of the 10-K-trained algorithm given the business-oriented nature of its articles. Next, for each topic in each rolling window, we randomly select three of the 10 most probable words associated with the topic based on the word weights produced by the LDA model.[33] An intruder word is selected at random from the pool of top 10 words appearing in another random topic.[34] We then use the word embedding algorithm to test all possible word combinations across all topics discovered across the rolling windows in our sample (i.e., three topic words plus one intruder word, which amounts to roughly 12.5 million possible combinations).

Our results show (not tabulated) that the algorithm correctly identifies the intruder words

[32]We use the word2vec algorithm created by Google to execute our word intrusion task (Mikolov et al. [2013]). The algorithm essentially uses a neural network to translate the relative meanings of words into measurable distances. As an illustration, the algorithm would be able to reconstruct country capital cities by using the equation, $vec(\text{`Paris'})vec(\text{`France'}) + vec(\text{`Italy'}) = vec(\text{`Rome'})$.

[33]Chang et al. [2009] and Dyer, Lang, and Stice-Lawrence [2017] select the five most probable words in a given topic. We select a smaller number of words to make our human subjects task more feasible given that our experiment is more comprehensive than the process used in the two studies. For instance, Chang et al. [2009] present 10 word evaluation tasks to 8 subjects, whereas our experiment presents 20 evaluation tasks to 180 subjects. To test whether our results are robust to this design choice, we replicate our machine-based word intrusion task using the five most probable topic words plus one intruder word. The inferences from this test is quantitatively similar to those based on our small word set.

[34]The pool of unrelated words excludes overlapping words that rank within the ten most probable words in both topics. This adjustment is necessary because, as noted earlier, the LDA model allows the same word to be associated with multiple topics.

with an average accuracy of 53% (50%) when trained using the corpus of 10-K filings (*WSJ* articles).[35] A one-tailed t-test confirms that this precision rate is statistically greater than random chance (25% or 1 out of 4 words) at the 1% level. This level of topic coherence compares well to an objective word embedding task of matching countries with country capitals using both a general Internet corpus and the set of WSJ articles (average accuracy of 57% and 58%, respectively).

Similar to Chang et al. [2009], we use the online labor market of Amazon Mechanical Turk (MTurk) to conduct our human-subject word intrusion task.[36] Our MTurk tasks are constructed similarly to the tasks performed by the word embedding algorithm. We however make the procedure practicable by presenting each participant with a small subset of the word sets evaluated by the machine algorithm. Specifically, once participants access our online experimental materials, they are each presented with 20 word sets that are selected at random from a pool of 200 word sets. The 200 word sets are randomly selected from the full set of word combinations evaluated by the algorithm. Each participant must then identify the intruder word for each word set.

Our subject pool consists of 180 participants who are U.S. citizens, proficient in English, and are at least 18 years of age.[37] We collect demographic information at the end of the task to assess participants familiarity with financial terminology. Our participants are on average 34 years old and about 85% have completed some college and above. Sixty-four percent of the participants report using financial statements at least once to evaluate a firm's performance.

[35]For additional robustness, we also use Stanford NLP's GloVe model trained on a general Internet corpus of 840 billion words (Pennington, Socher, and Manning [2014]). The average accuracy for the Internet-trained algorithm is comparable at 49%. Appendix A.4 of the online appendix provides further discussion on the results from our machine-based word intrusion tests.

[36]The Institutional Review Board (IRB) at the researchers' universities approved the use of human subjects for the experiment reported in this paper.

[37]We limit participation to U.S. citizens due to policies at the researchers' universities that impose restrictions on payments to non-U.S. citizens. We determine English proficiency by asking participants to indicate whether English is their native language. Our original sample consists of 200 participants. We eliminate 12 responses with duplicate IP addresses and 2 responses that indicate English as a non-native language. We also exclude 6 participants who completed the task in less than 60 seconds to minimize concerns about participant effort in online labor markets (Farrell, Grenier, and Leiby [2017]). Our results however do not change when we include all 200 responses.

Lastly, 60% of the subject pool have bought or sold a company's common stock or debt securities, while 67% plan on investing in a company's securities within the next five years. Overall, these demographics suggest that our MTurk participants have reasonable levels of financial knowledge, consistent with the evidence reported in Farrell, Grenier, and Leiby [2017].

The results from the experimental task indicates that participants correctly identify the intruder word about 40% of the time. This precision rate is statistically higher than random chance (one-tailed t-statistic = 14.72). Although the human precision rate is lower than that achieved by the machine algorithm, it still compares well when we consider that the word embedding algorithm was first trained on the *entire* corpus of 10-K narratives.[38] We however acknowledge that the lower rate could reflect noise in the topics derived from the LDA model. In summary, our evaluation methods suggest that the LDA algorithm provides a valid set of semantically meaningful topics that are reasonably coherent and interpretable by human judges. These results provide us with greater confidence for investigating the prediction accuracy of *topic* relative to quantitative financial statement and stock market variables (RQ1) and textual style characteristics (RQ2).

4.2 Predictive Value of LDA Topic Measure

4.2.1 Empirical Methodology

To investigate our research questions, we first estimate in-sample prediction models using rolling five-year windows. 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 subsequent to the end of each rolling window.[39] [40] We begin our analyses by estimating

[38] We do not conduct a comparative word intrusion task such as the objective country-capital city task given our inability to control MTurk participants' access to outside resources.

[39] For filings coded as an irregularity restatement, we ensure that the restatement is revealed by the end of the in-sample window. We are unable to apply this restriction in the AAER sample as the UC Berkeley dataset does not include the release dates of the AAERs.

[40] For example, the estimated results for 1994 – 1998 (1995 – 1999) are used to predict misreporting for a holdout sample of firms in 1999 (2000) and so on.

logistic regressions of *misreport* on vectors of the disaggregated topic proportions (*topic*) as follows:

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

We estimate equation (1) for the five-year window preceding each of the prediction years, 1999 to 2012. As noted earlier, we lack sufficient AAER events for the 2011 and 2012 years, and thus exclude these two years when conducting our out-of-sample AAER tests. Similar to Dechow et al. [2011], we construct a prediction score (*p_misreport*) using the estimated coefficients and apply this scoring in our out-of-sample tests as follows:

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

We estimate two additional regression specifications to examine RQ1. The first specification replaces the *topic* vector with the vector of financial variables (*F-score*) outlined above. The second specification extends equation (1) by including both vectors of *topic* and the financial variables.[41] In both cases, we generate *p_misreport* and run the out-of-sample tests. For RQ2, we introduce four specifications that include our textual style measures (denoted *Style*). The first model includes our textual style metrics with the second including both *topic* and *Style*. Lastly, we introduce a comprehensive model that jointly tests the incremental predictability of topic over both financial metrics and textual features. This model expands equation (1) by including vectors of *F-score* and *Style*. For completeness, we also estimate an alternative model with *F-score* and *Style*, i.e., before adding *topic* as an additional set of predictors. Our general regression form for the comprehensive model is

[41] Our restructuring indicator variable is valid only for the 2000 fiscal year and onwards due to the lack of restructuring data in Compustat for prior years. We therefore exclude the restructuring variable from our estimations of equation (3) for five-year windows that do not overlap the 2000 fiscal year.

specified below in equation (3):

$$\log\left(\frac{misreport_{i,t}}{1-misreport_{i,t}}\right) = \alpha + \sum_{j=1}^{17}\beta_j F\text{-}score_{j,i,t} + \sum_{j=1}^{20}\beta_{j+10} Style_{j,i,t} + \sum_{j=1}^{20}\beta_{j+40} topic_{j,i,t} + \varepsilon_{i,t},\ t \in [T-5, T-1],\ i \in \text{Companies} \tag{3}$$

We tightly control the convergence of our logistic regressions given the large number of predictors and the low frequency of AAERs and irregularity restatements in our test windows (see Table 1). We control the convergence by conducting checks for both completeness and quasi-completeness of each regression specification. Appendix A.3 of the online appendix details the necessary steps for conducting these checks.

Given the structure of our rolling time-series analysis, we are unable to use a standard Fama-MacBeth methodology to pool our results for the predicted window. This restriction results from the time-varying nature of *topic* as previously reported. Thus, we cannot aggregate across on a variable level. To address this research design issue, we use Fisher's (1932) method to provide aggregated test statistics.[42] The Fisher test statistic is appropriate for our analyses since the out-of-sample regressions are estimated using non-overlapping years. We refine our test statistic further by deriving a statistic referred to as a Var-Gamma test (see Appendix D). This test statistic allows us to compare the results of Fisher's method, statistically testing whether one detection model performs better than another when pooled across years.

To further aid the interpretation of our results, we follow prior studies (e.g., Kim and Skinner [2012], Larcker and Zakolyukina [2012], Gutierrez et al. [2017], Bao et al. [2018]) and present a general measure of each model's out-of-sample classification performance using the area under the receiver operating characteristics (ROC) curve. The ROC curve is a two-dimensional plot across different cut-off 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

[42]The test statistic is computed as $-2\sum_{i=1}^{N}\log(p_i) \sim X^2_{2N}$, where p_i is the ith p-value of N total p-values.

curve (AUC) is a widely-used indicator of a model's predictive ability since 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, in essence, represents the probability that a randomly chosen misstated year will be ranked higher by the respective model compared to a randomly chosen non-misstated year.

As Bao et al. [2018] point out, any reasonable fraud detection model must have an AUC that is higher than that of a random classification model (i.e., the AUC should be greater than 0.50). We therefore use this threshold as one of our performance benchmarks when evaluating our AUC results. To compare the out-of-sample AUCs across our predictive models, we use the methodology developed by Janes, Longton, and Pepe [2009] and conduct bootstrapped, non-parametric Wald tests (based on 1,000 bootstrapped replications) of the differences in the AUC statistics. Given our rolling time-series analyses, we report pooled AUC statistics and p-values for tests of the AUC differences between our prediction models. Specifically, we compute the AUCs and Wald test p-values using pooled data for all sample years with bootstrapped standard errors corrected for clustering by year. To assess whether the pooled AUC statistic for each regression specification is significantly greater than 0.50, we compare the statistic against the AUCs produced using simulated random data bootstrapped with 1,000 replications. [43]

4.2.2 In-sample Predictive Value of *topic*

Before turning to our primary research questions, we conduct in-sample tests to evaluate whether topic performs reasonably well in detecting misstatements. Our AAER sample consists of 29,785 total observations across all prediction years for the 1999 to 2010 period, while the restatement sample consists of 34,293 observations over the 1999 to 2012 period

[43]Following Gutierrez et al. [2017], we conduct our AUC tests using the rocreg Stata command. The command does not assume that the ROCs are independent and calculates standard errors based on bootstrap and clustering by year. We use pooled data for these tests since this approach is superior to taking a simple average of the AUCs across each test year.

(see Table 1 for annual counts for each misreporting sample). Figures 1 and 2 depict the distribution of each combined topic over our misreporting prediction years and whether the topic is significantly associated with financial misreporting as proxied by the occurrence of an irregularity restatement (Figure 1) or a SEC enforcement action (Figure 2). Again, for ease of interpretation, we present aggregated results based on the combined topics across each estimation year and report industry-specific topics at the bottom of each figure. We determine the significance of the combined topics by estimating yearly in-sample regressions of the disaggregated subtopics (i.e., the topics associated with a given combined topic in each year) on our misreport indicator variable. We orthogonalize the subtopic proportions to 2-digit SIC industries to control for unobserved industry effects.[44]

We observe in both 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 post-retirement benefits (3), fuel costs and purchase commitments (4), digital technology and services (15), and industry-specific topics such as aircraft leasing arrangements (50). Other topics appear later in the sample period, indicating the evolving nature of firms' management communications. For instance, discussions of collaborative business arrangements such as joint ventures (27), strategic alliances (32), and partnerships (37) are more prominent in the second half of our prediction period. Likewise, discussions of securitized/guaranteed securities (41) appear prominently in 2008, coinciding with the turmoil surrounding asset-backed securities markets during that time period.

With respect to the ability to detect misreporting, Figure 1 illustrates that discussions of increases in income performance compared to prior periods (combined topic 2) is significantly associated with irregularity restatements in all but one of our prediction years. However, the direction of the significance is not consistent throughout the sample period. We also observe that discussions of declines in income performance (topic 1) is significant in relatively few years in our sample. These results suggest that the association between misreporting and

[44]Our results are qualitatively similar when we additionally orthogonalize by audit firm type (Big N, mid-size, or small) to control for unobserved audit firm effects.

managerial discussion of financial performance is not as clear cut as suggested by prior work linking poor financial performance to accounting manipulations.

The results in Figure 2 also suggest that misreporting firms are more likely to discuss issues related to advertising expenditures, investments in securitized/guaranteed securities, strategic alliances, and growth in franchised operations. Combined topics that load consistently negative include discussions of merger activities, foreign country risks, hedging activities, legal proceedings, credit arrangements, and stock option plans, suggesting that restatement firms are less likely to discuss these issues in misstatement years. The results for AAER firms (Figure ?) are similar, but with some variation in the timing of the topic loadings. For instance, AAER firms are less likely to discuss merger activities and increases in income performance, especially in early sample years.

4.2.3 Predictive Value of *topic* versus Financial Variables (RQ1)

Table 2 presents separate summary statistics of our financial variables for misstated and non-misstated firm-years in the AAER and irregularity restatement samples. Some of the variables behave similarly across the two samples. For instance, both samples show that misreporting firms are more likely to issue securities and engage in restructuring activities during misstatement years. Several variables, however, show opposing differences between the two samples, with many failing to show significant differences in the irregularity restatement sample. While these opposing results likely reflect the more egregious nature of AAERs as well as potential selection issues, they highlight the difficulty in establishing clear associations between misreporting and quantitative financial statement and stock market measures (Dechow et al. [2011], Purda and Skillicorn [2015]).

Table 3 presents out-of-sample tests of the predictive role of *topic* and our vector of financial variables (denoted as *F-score*). Panels A and C present Fisher test statistics and pooled AUCs for AAERs and irregularity restatements, respectively. The Fisher statistics indicate that the financial variables (*F-score*) provide a significant amount of information for

predicting AAERs ($p = 0.000$). However, they fail to provide significant informational value for predicting irregularity restatements ($p = 0.165$), consistent with the summary statistics reported in Table 1. The AUC statistics provide similar evidence. The AUC for the *F-score* model is markedly higher for the AAER sample, with a gain over a random classification model of 24% compared to a gain of 9% for the irregularity restatement sample (AUC of 0.742 versus 0.589, respectively; both AUCs are statistically greater than 0.50 at the 1% level using our simulated bootstrap test procedure).

Panels B and D present Var-Gamma and Wald tests of the predictive ability of *topic* compared to that of *F-score* and a joint specification of both predictors. The Var-Gamma (Wald) test evaluates the statistical significance of the difference in the Fisher (AUC) statistics across pairs of prediction models. In the case of AAERs (see Panel B), the results from both performance measures indicate that the stand-alone *topic* vector underperforms the vector of financial metrics (Var-Gamma $p = 0.000$ and Wald $p = 0.057$). For instance, the pooled AUC for the stand-alone *topic* vector is slightly lower at 0.700 versus 0.742 for the *F-score* vector. These results are not surprising given that the original Dechow et al. [2011] *F-score* model was built using a sample of AAERs. Nonetheless, we find that the paired vectors of *topic* and *F-score* perform significantly better at predicting AAERs than the stand-alone *F-score* vector (Var-Gamma $p = 0.000$ and Wald $p = 0.002$). Specifically, the AUC comparisons indicate that the addition of our *topic* measure to a standard *F-score* model increases predictive accuracy by 2.7%.

With respect to irregularity restatements (see Panel D), we find that the stand-alone *topic* vector and the pairing of *topic* with *F-score* both outperform the stand-alone *F-score* specification. For instance, the AUC statistics show that the stand-alone *topic* vector improves predictive accuracy by 2.7% when compared to the stand-alone *F-score* model (AUC of 0.616 versus 0.589, Wald $p = 0.065$). Detection ability is even greater (increase in AUC of 4.1%, Wald $p = 0.000$) when we compare the performance of a joint model of *topic* and *F-score* with that of the stand-alone *F-score* model. In comparing the predictive ability of

topic with that of *topic* paired with *F-score*, the Var-Gamma statistical test is insignificant ($p = 0.355$), while the difference in the AUCs is negative and statistically significant (decline in AUC of 0.014, Wald $p = 0.075$). These results suggest that financial metrics add little detection power over *topic* for the restatement sample. In other words, the stand-alone *topic* model achieves a level of predictive accuracy that is fairly equivalent to a joint *topic* and *F-score* model. Overall, our evidence from Table 6 indicates that *topic* contributes significant incremental power in identifying misreporting events beyond traditional financial metrics. More importantly, our evidence suggests that *topic* serves as a better detection tool when assessing the likelihood of irregularity restatements, while a joint model of *topic* and financial metrics performs best when detecting more egregious misstatements involving AAERs.

4.2.4 The Predictive Value of *topic* versus Textual Style (RQ2)

Table 4 presents separate univariate statistics for our style characteristics for misstated and non-misstated firm-years. We find that many of the style features show inconsistent differences across both samples and in some cases, show differences that contradict conventional notions. For instance, misstated filings in the irregularity restatement sample have more bulleted (and presumably, more concise) information, lower lexical variety, and more active voice grammar relative to non-misstated filings. Misstated filings in the AAER sample exhibit slightly shorter sentences, less complex paragraphs, less positive tone, and greater readability (see Fog index). These opposing results are not unique to our study as Purda and Skillicorn [2015] find similar contradictory evidence of more deceptive and negative language in filings classified as "truthful." Such evidence underscores the potential pitfalls in relying on basic textual measures to identify financial misreporting.

We approach RQ2 by combining the *topic* and textual style (*Style*) vectors in the same regression model. Table 5 presents the Fisher and pooled AUC statistics for out-of-sample tests of the predictive performance of *topic* relative to *Style*. Panels A and B presents the test statistics for AAERs; Panels C and D presents the results for irregularity restatements.

The evidence in Panels A and C suggests that a joint model of *topic* and *Style* is a good predictor of misstatements involving AAERs and irregularity restatements (e.g., AUCs of 0.735 and 0.669, respectively; both AUCs are statistically greater than 0.50 at the 1% level). In the AAER sample, the performance measures in Panel B indicate that *topic* by itself is a better or at least an equivalent predictor of misreporting than the stand-alone *Style* model (Var-gamma $p = 0.000$ and Wald $p = 0.180$). A combined prediction model of *topic* and *Style* also outperforms the stand-alone *Style* model in the AAER sample (Var-Gamma $p = 0.000$ and Wald $p = 0.000$).

The performance measures for irregularity restatements (see Panel D) show that, while *Style* is a better predictor than *topic* (e.g., AUC of 0.663 versus 0.616, Wald $p = 0.000$), the joint model of *topic* and *Style* is a better predictor than the stand-alone *topic* and *Style* models.[45] Thus, we find that the best specification for predicting AAERs is *topic* by itself, while the best specification for predicting irregularity restatements is *topic* paired with *Style*. This evidence suggests that detection models based on *topic* and textual style characteristics are better able to identify accounting manipulations that do not involve a SEC enforcement action.

4.2.5 Joint Predictive Value of *topic*, Financial Variables, and Textual Style

We conduct extended analyses of the interplay between all three sets of prediction variables: *topic*, financial statement and stock market variables (*F-Score*), and textual style characteristics (*Style*). This comprehensive analysis evaluates whether the predictive ability of *topic* is robust to the inclusion of both financial and textual style characteristics. Table 6 presents out-of-sample results from a comprehensive regression of all three vectors of prediction variables. For brevity, we do not present benchmark performance measures for all model combinations of the *topic*, *F-score*, and *Style* vectors. However, for the AAER sample

[45]The Var-Gamma test statistics indicate that the joint model of *topic* and *Style* dominates stand-alone models of either *topic* or *Style*. The AUC comparisons lead to a similar conclusion, except that the joint *topic* and *Style* model outperforms the stand-alone *topic* model but performs the same as the stand-alone *Style* model.

(see Panels A and B), we present performance statistics for the joint model of *topic* and *F-score* since the results in Table 3 indicate that this model performed well in predicting AAERs. Likewise, we present performance statistics for the joint *topic* and *Style* model for the irregularity restatement sample (see Panels C and D) since this model dominates the stand-alone *topic* and *Style* models in Table 5.

In Panels A and C of Table 6, we find that the three-vector specification performs reasonably well in detecting accounting misstatements out of sample. The AUC statistics for the AAER and irregularity restatement samples are well above the 0.50 threshold (AUCs of 0.778 and 0.670, respectively). The results in Panels B and D indicate that this joint model performs better than the stand-alone *topic* model in predicting both types of accounting misstatements (the Var-Gamma and Wald statistics for the model comparisons are all significant at the 1% level). However, the combined vector does not perform any better than the joint model of *topic* and *F-score* in predicting AAERs (Var-Gamma $p = 0.393$ and Wald $p = 0.331$ in Panel B) nor the joint model of *topic* and *Style* in predicting irregularity restatements (Var-Gamma $p = 0.684$ and Wald $p = 0.744$ in Panel D). This evidence corroborates our previous results—*topic* and quantitative financial measures are strong predictors of misstatements involving AAERs, whereas *topic* and aggregate textual measures provides more robust power for predicting irregularity restatements.

4.3 Economic Significance of LDA Topic Measure

To gauge the economic significance of our results, we examine the classification accuracy of the out-of-sample models using cut-offs 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 percentile as "high risk." These accuracy rates are equivalent to the true positive rate or sensitivity of each predictive model at various cut-offs. We compute these rates for each prediction year and report the average annual percentage of misstated 10-K filings that are accurately classified as misstated

by each model at the respective percentile cut-offs.

We also follow Bao et al. [2018] 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 predicted scores. The NDCG@k is a measure of ranking quality and has been shown to be theoretically consistent in evaluating classification accuracy across models (Wang et al. [2013]).[46] The 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 a cut-off top position or percentile k (denoted DCG@k). The position value k represents the number of predicted scores over which the model's ranking quality is evaluated. We set k at 1% so that the number of predicted scores in the ranking list represents the 99th percentile of the test sample in a given prediction year. This k value of 1% is consistent with the average frequency of misreporting events over our sample period as reported in Table 1. The DCG@k measure is further normalized to allow for comparisons of ranking quality across multiple models. This normalized measure (NDCG@k) is computed by scaling DCG@k by its ideal value assuming a perfect prediction model (i.e., the DCG@k for an ideal model where all true misreporting events are ranked at the top). This normalization leads to a NDCG@k measure that ranges from 0 to 1, with a higher value indicating better ranking performance of a given prediction model.

Table 7 reports the percentage of filings that are correctly classified as misstated using multiple models for the AAER (Panel A) and irregularity restatement samples (Panel B). From Panel A, the results for the 50th percentile cut-off indicates that *topic* by itself captures roughly 72.54% of misstatements involving AAERs. The joint model of *topic* and *F-Score*, or alternatively *topic* and *Style*, performs slightly better flagging about 74% of misreported

[46]The NDCG@k measure is formally defined as DCG@k scaled by Ideal DCG@k, where DCG@k = $\Sigma_{i=1}^{k} \left(2^{rel_i} - 1\right) / \log_2 (i+1)$. The position value k is equal to 1% and rel_i equals 1 if the i-th observation in the ranking list is a true misreporting event, and 0 otherwise.

filings, while the three-vector model correctly flags 75.09% of misstatements. This evidence corroborates our findings in Table 6. When we focus on high-risk prediction scores, we observe that the three-vector model and the joint model of *topic* and *F-score* capture the most misstatements at the 90th percentile (31.50% and 32.07%, respectively). These accuracy rates are roughly 33% higher than the rate achieved using a standard model of *F-score* paired with *Style* (23.73%). At the 95th percentile, the three-vector model once again dominates with an accuracy rate that is 46% higher than that for the joint *F-score* and *Style* model (21.44% versus 14.66%). The NDCG@k measure at the 99th percentile indicates that the joint model of *topic* and *F-Score* model has the highest ranking performance at 0.192. This value is about 5 (11) percentage points higher than the stand-alone *F-score* (*Style*) model and 2.4 percentage points higher than the joint *F-score* and *Style* model.

In Panel B, the three-vector model and the joint *topic* and *Style* model are the most efficient at capturing irregularity restatements at the 50th percentile cut-off. However, these models outperform their benchmark models (the joint *F-score* and *Style* model and the stand-alone *Style* model, respectively) by less than 1 percentage point. For high-risk observations, the three-vector model is most accurate, detecting roughly 28.05% and 16.83% of misstated filings at the 90th and 95th percentile cut-offs, respectively. When we compare the three-vector model against the joint *F-score* and *Style* model, we find that including *topic* as an additional predictor increases detection accuracy by roughly 5-6% at both percentile cut-offs. The NDCG@k measure portrays a similar picture for prediction scores in the 99th percentile. Specifically, adding *topic* to the benchmark *F-score* and *Style* model increases classification accuracy by roughly 2 percentage points. Taken together, our results in Panels A and B suggest that *topic* is an economically significant predictor of financial misreporting and that the incremental value of *topic* is especially salient when attempting to identify high risk accounting practices.

To further illustrate the economic significance of our results, we follow prior studies (e.g., Dechow et al. [2011]) and assess 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 taken into account. 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, i.e., the misstated filings pertaining to the 1998 and 1999 fiscal periods, respectively.[47]

Our detection model classifies Enron's 1999 filing as misstated based on a prediction score that ranks at the 93rd percentile across all filings issued in 1999. Of all the model's predictors, two variables contribute the most to Enron's prediction score. The first variable 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. Turning to the 2000 calendar year, we find that Enron's 10-K filing is classified as misstated at an even higher percentile (98.5). The results also show that Enron's discussion of income increases (combined topic 2) is substantially lower in the 2000 filing, ranking in just the 2nd percentile relative to industry peers. This dramatic shift could reflect deliberate disclosure decisions by Enron executives to divert attention away from soaring earnings and the sources of its revenue growth.[48] [49]

[47]The enforcement actions related to Enron cite material accounting violations for the fiscal years, 1997 to 2000 (see SEC AAER No. 1640 and 1821).

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

[49]We also conduct a second case study of the SEC enforcement action filed against Zale Corporation for the improper capitalization of television advertising costs over the 2004 to 2009 period. We find that our prediction model correctly classifies Zale's 10-K filings as misreported at the 97th percentile and above in all years except 2004. The topics that contribute the most to Zale's prediction score point to high amounts of discussion related to media and entertainment (combined topic 20), and digital technology and services (combined topic 15).

4.4 Extended Analyses

4.4.1 Predicting the first instance of misreporting

Our main analysis requires a misreporting event to be publicly known by the end of the in-sample estimation window for it to be included in that window (see discussion in footnote 39). There are no cases in our sample where the same misreporting event affects multiple annual reports spanning both the in-sample estimation period and the out-of-sample prediction period. 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 period and another appearing in the prediction period). While such cases are due to unrelated misreporting events, their inclusion raises the concern that our *topic* measure could be biased towards identifying repeat offenders or certain types of firms, rather than detecting variations in thematic content when firms engage in misreporting. To address this concern, we adjust our misstatement samples by reclassifying *misreport* to be 1 only if *misreport* was 0 for the same firm in the previous year. This adjustment creates an out-of-sample dependent variable that only picks up the first year affected by a misstatement and ignores any repeated misstatements unless there is at least 1 non-misstated year in between the events. This test is empirically challenging given the rarity of misreporting events. In fact, in our out-of-sample prediction years, there are only 100 first instance AAER events (down from 380 AAER firm-years) and 520 first instance irregularity restatements (down from 648 irregularity restatement firm-years). As such, we expect our results to be noisier, but fairly consistent with our main results.

In untabulated results, we find that the joint *topic* and *F-Score* model performs best for the AAER sample with the highest Fisher statistic of 53.52 ($p = 0.000$) and the highest AUC of 0.682 (significantly greater than 0.50 at the 1% level). Our comparative tests of predictive performance indicate that the joint *topic* and *F-score* model performs better in detecting AAERs when benchmarked against all other specifications, except for the stand-

alone *F-score* model which performs equivalently when we compare the Fisher and AUC statistics (Var-Gamma $p = 0.225$ and Wald $p = 0.307$). When we compare the classification accuracy of each model, we find that the incremental value of *topic* is significantly stronger in detecting the first instance of AAER events at the 90th percentile cut-off and higher. In such cases, the three-vector model and the joint *topic* and *F-score* model outperform all other specifications.

Turning to the irregularity restatement sample, we find that the joint *topic* and *F-score* model is the only specification with a statistically significant Fisher statistic (Fisher statistic of 49.53, $p = 0.004$). The AUCs for all our specifications are statistically higher than 0.50, with the stand-alone *Style* model achieving the highest AUC of 0.596. In terms of comparative performance, the joint *topic* and *F-score* model again outperforms all other models, including the *F-score* model, when we examine differences in the AUCs and the Fisher statistics. The *Style* specification also outperforms the *F-score* model when detecting irregularity restatements (increase in AUC of 2.7%, p = 0.044), but performs similarly when benchmarked to the joint *topic* and *F-score* model. Overall, it appears that *topic* improves the out-of-sample prediction of the first instance of an AAER event for a given firm-year, primarily when detecting the first instance of high-risk AAER events. However, it appears that *topic* and *Style* are less complementary for detecting the first instance of an irregularity restatement, relative to our previously reported results based on the full set of irregularity restatements (see Tables 5 and 6).

4.4.2 Alternative out-of-sample measurement

Our next set of analyses (not tabulated) impose additional out-of-sample restrictions to further control for potential biases arising from the inclusion of firms with repeated but unrelated instances of misreporting in both the estimation and prediction periods. These restrictions lead to prediction periods that are closer to being out-of-sample with respect to a given firm. We must however caution that the removal of firms with repeated misreporting

events is likely to create a look-ahead bias since it is unknown in a given prediction year whether the same firm will engage in future accounting irregularities. Nonetheless, our results from these tighter hold-out samples continue to demonstrate the strong incremental power of topic in detecting misreporting.

We impose additional restrictions on the irregularity restatement sample only given the high rarity of AAERs over our prediction windows.[50] Our first restriction removes all sample firms for which *misreport* is set to 1 in both the in-sample estimation period and the out-of-sample prediction period in each rolling window. This adjustment removes 67 repeat offenders from our irregularity restatement sample across all prediction windows. The fact that we remove only 67 restatement events (9.6% of our entire sample of irregularity restatements) indicates that biases due to repeat offenders is unlikely to be a major problem. Indeed, the results from this restricted hold-out sample are similar to our main results and show that the model specification of *topic* paired with *Style* performs best in predicting irregularity restatements.

The second restriction removes all firms for which misreport is set to 1 in both the estimation and prediction period in any window. In other words, we remove these firms entirely from our sample. This restriction removes 53 repeat restatement offenders from our test sample. The replicated results using this tighter sample indicate that the inclusion of the *topic* vector improves detection performance when benchmarked to a stand-alone *F-score* model. We also find that the three-vector model provides the strongest detection power, though its performance is only marginally higher than the standard *F-score* and *Style* model. In sum, the combined results indicate that our inferences are robust to additional hold-out sample adjustments for firms with multiple misreporting events.

[50]We lose about 60% of the AAERs in our prediction windows when we impose these restrictions. This is because SEC enforcement actions are more likely to involve multiple filing years. While the additional restrictions significantly lowers our sample power, we continue to find that a joint model of *topic* and *F-score* performs best when we use the restricted hold-out samples of AAERs. This result is consistent with the inferences drawn from our results in Tables 3 and 5.

4.4.3 Predicting long-duration misreporting events

We conduct further analyses to investigate whether our *topic* construct improves the detection of not just the first instance of misreporting but also misstatements affecting multiple annual reports over consecutive years. This evidence is warranted since accounting manipulations can go undetected for long durations of time (Singer and Zhang [2018]).[51] In fact, our sample of irregularity restatements (AAERs) include accounting manipulations that affect up to five (six) consecutive annual reports. We perform our tests by examining the out-of-sample classification accuracy of our models when detecting the first instance as well as consecutive instances of misreporting. Similar to our analysis in Table 7, we compute classification rates for our prediction models based on the number of consecutive misstated filings and at various percentile cut-offs. These classification rates inform us of the incremental value of topic in predicting misreporting events with durations longer than one year. We illustrate the incremental classification accuracy of *topic* by misreporting duration in Appendix A.5 of the online appendix.

Consistent with our analysis of the first instance of misreporting, we find that *topic* adds significant incremental predictive power when added to a joint model of *F-score* and *Style*. As discussed earlier, this result holds primarily for detecting the first instance of high-risk AAERs (i.e., those at the 95th percentile and above). The *topic* vector greatly improves classification accuracy at all three cut-offs when detecting AAERs involving consecutive misreporting years (i.e., AAERs affecting two or more annual filings). The incremental power of *topic* is less salient but still economically significant when detecting consecutive years of irregularity restatements. Specifically, the three-vector specification outperforms or is equal to the joint *F-score* and *Style* model at all three cut-offs over all durations of irregularity restatements. Taken together, our results indicate that *topic* provides economically significant value when assessing the likelihood of misreporting, irrespective of the duration of the

[51] Using a sample of irregularity restatements over the 2000 to 2013 period, Singer and Zhang [2018] find an average duration of close to 2 years for misstatements that affect consecutive annual filings.

misreporting.

5 Robustness Tests

In this section, we conduct a series of sensitivity checks for our primary results. Our first test examines the usefulness of *topic* in detecting restatements attributable to unintentional misapplications of GAAP (i.e., purely accounting errors). Next, we re-estimate *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, replicate our analyses using the raw *topic* proportions of each filing (as opposed to the normalized *topic* proportions). We do not tabulate these results for the sake of brevity but briefly discuss each sensitivity check below.

5.1 Annual differences in the frequency of misreporting events

As noted earlier, the frequency of AAERs declines substantially after 2005, presumably due to the length of time necessary for the SEC to investigate and process an enforcement action. We also observe that the frequency of irregularity restatements is more stable in the post-2002 period following the 2002 Sarbanes-Oxley Act. To ensure that our results reflect a complete set of true misreporting events, we limit the prediction years for the AAER (irregularity restatement) sample to the 1999 to 2005 (2003 to 2012) period. We then replicate our detection models using these alternative test windows. In untabulated results, we continue to find that the joint model of *topic* and *F-score* performs best in detecting AAER events over the 1999 to 2005 period, with the three-vector model performing equivalently well. The three-vector model and the joint model of *topic* and *Style* also dominates in detecting irregularity restatements occurring over the 2003 to 2012 period.

5.2 Unintentional Misstatements

We investigate our models' ability to detect restatements involving unintentional errors (i.e., misstatements stemming from accounting mistakes and data errors). Since errors do not reflect explicit intent to report misleading information, we expect the incremental value of *topic* to be lower in this setting. Consistent with this notion, we find that, while all specifications perform reasonably well in predicting unintentional accounting errors, the incremental value of our *topic* measure is economically smaller relative to our reported results. Nonetheless, given that some errors can escalate to intentional manipulations (perhaps to conceal the error), our evidence suggests that *topic* provides at least some added value as an early warning sign.

5.3 Using MD&A Text

Several text-based studies of financial misreporting examine the MD&A section of the 10-K filing (see e.g., Cecchini et al. [2010a], Hoberg and Lewis [2017], Purda and Skillicorn [2015]). We therefore investigate whether our results differ when we restrict our textual analysis to the MD&A section. We reconstruct our *topic* and *Style* variables using the text extracted from the MD&A section (see Appendix A.1 of the online appendix for further details). The out-of-sample results show weaker predictive performance compared to our reported results. However, *topic* continues to provide significant incremental power in detecting misreporting, especially in the case of irregularity restatements. Overall, these results reaffirm prior arguments that the entire 10-K filing provides additional content that is useful for drawing inferences in text-based research (see Li [2010b], Loughran and McDonald [2016]).

5.4 Regularized Logit Regression

Due to the relative rarity of misreporting events, there is some concern of overfitting given the large number of predictors in our models (see Perols et al. [2017]). We address this issue by

using an L1 regularized logistic regression to re-estimate the results. The L1 regularization approach applies a penalty for increasing the number of independent variables, thereby controlling for biases arising from model overfitting. All of our out-of-sample results are similar to our main analyses with one exception—the joint specification of *topic* and *F-Score* becomes the sole best predictor for AAERs. As such, the full three-vector specification fails to overcome the penalties applied by the L1 estimation process.

5.5 Raw *topic* Measure

Our final sensitivity check uses the raw *topic* proportions instead of the normalized proportions. This approach increases the variance of *topic*, as the measure is now influenced by the amount of text in the document. The prediction results for the AAER and irregularity samples are qualitatively similar to our reported results. The raw *topic* measure performs slightly worse in our comparative tests, but the incremental predictability of *topic* remains strong in general. We therefore conclude that the amount of each topic, rather than the proportion, is also useful for detecting accounting misstatements.

6 Conclusion

In this study, we employ a sophisticated textual analytic tool to directly detect *what* is being disclosed in 10-K filings (as opposed to *how* it is being disclosed). More specifically, we develop a unique measure, labeled as *topic*, which simultaneously identifies and quantifies the thematic content of annual financial statements. Drawing on the communications and management disclosure literature, we assess whether this *topic* measure is incrementally informative in predicting intentional misreporting compared to standard quantitative financial measures and textual style features.

Using SEC AAERs and irregularity restatements to identify misstated filings, we find that our *topic* measure provides significant incremental predictive power over commonly-

used financial statement and textual style measures. Specifically, based on out-of-sample tests, we find that models that incorporate *topic* outperform models based on financial and textual measures. Further, our results reveal that *topic* is incrementally and economically valuable in detecting above normal and high risk accounting misstatements, improving the prediction accuracy by as much as 46% in the case of SEC AAERs. We document that these results are robust to alternative model specifications, *topic* definitions, and the use of MD&A disclosures.

Our study makes several important contributions to the literature. First, we contribute to the financial misreporting literature by providing evidence that the thematic content of annual report narratives is useful in predicting intentional financial misreporting beyond traditionally examined financial measures and textual style characteristics. Second, we expand the burgeoning research in accounting that examines the textual portion of corporate disclosures. Specifically, we employ a robust textual analysis methodology, termed LDA, that quantifies the semantic meaning or thematic structure of financial statement disclosures. This approach is a significant step forward as it allows for broader insights into the correlation between language-based communication and financial misreporting. Our work also has significant practical implications for regulators and practitioners by demonstrating the usefulness of topic analysis in detecting high risk accounting practices. We hope this work spurs further discussion of the importance of non-numerical management disclosures, the information they contain, and their effects on capital markets.

Appendix A Identification of Irregularity Restatements

We conduct an automated text search of amended 10-K filings to identify irregularity restatements arising from intentional GAAP violations. 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 restatement as an unintentional accounting error. The search phrases for each classification criterion are laid out below. The '*' symbol indicates truncated words, while '...' 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 ...', '... former independent auditors ...', '... forensic or other account* ...', '... retain* ... independent legal counsel ...'

Appendix B Variables

Panel A: Financial Variables

Variable	Definition
$\log(TotalAssets)$	Log of total assets
$RSSTAccruals$	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]
$\Delta Receivables$	Change in accounts receivable scaled by average total assets
$\Delta Inventory$	Change in inventory scaled by average total assets
$\%SoftAssets$	Percent of total assets excluding PP&E and cash and cash equivalents
$\Delta CashSales$	Percentage change in cash sales, where cash sales is measured as total sales minus the change in accounts receivable
$\Delta ReturnOnAssets$	Change in income before tax, scaled by average total assets
$ActualIssuance$	An indicator coded as 1 if the firm issued debt or equity securities during the year, 0 otherwise
$OperatingLeases$	An indicator variable coded as 1 if future operating lease obligations are greater than zero, 0 otherwise
$Book\text{-}To\text{-}Market$	The ratio of total common equity to the market value of equity, where market value is computed as total common shares outstanding multiplied by the fiscal year end closing share price
$Lag(Mkt\text{-}AdjReturn)$	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
$BigNAuditor$	An indicator variable coded as 1 if the firm was audited by a Big N auditor in the current fiscal year, 0 otherwise.
$Mid-sizeauditor$	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 non-zero restructuring charges during the current fiscal year, 0 otherwise

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 included in the 10-K filing
$Tags$	The length of all HTML tags used in the 10-K filing
$ParsedSize$	The number of characters in the 10-K filing after parsing (see Appendix A.1 of the online appendix for the full parsing methodology)
$SentenceLength$	Mean sentence length, in words
$WordStddev$	Standard deviation of word length
$ParagraphStddev$	Standard deviation of paragraph length
$Repetitions$	The mean number of times each sentence is repeated in the parsed 10-K filing
$SentenceStddev$	Standard deviation of sentence length
$TypeTokenRatio$	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
$Coleman\text{-}LiauIndex$	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 (3 or more syllables) in the document
$\%ActiveVoice$	The percent of sentences written in active voice
$\%PassiveVoice$	The percent of sentences written in passive voice
$\%Negative$	The percent of negative words in the document based on the Loughran and Mcdonald (2011) dictionary of negative words
$\%Positive$	The percent of positive words in the document based on the Loughran and Mcdonald (2011) dictionary of positive words
$AllCaps$	The number of words in all capital letters with at least 2 letters
$ExclamationPoints$	The number of exclamation points in the parsed 10-K filing
$QuestionMarks$	The number of question marks in the parsed 10-K filing

Appendix C Combined Topics

1) Decrease in income compared to prior periods: compared to, gross profit, other income, company contributed, operating income, company expects, gross margin, income decreased, capital expenditures, decreased to
2) Increase in income compared to prior periods: compared with, gross margin, income was, operating income, gross profit, other income, fiscal compared, income taxes, non-interest income, profit was
3) Postretirement health care benefits assumptions: health care, care plans, assumed health, trend rates, care cost, cost trend, effect on the amounts, rates have, significant effect, postretirement health
4) Fuel costs and commitments: nuclear fuel, fuel related, coal mining, coal reserves, fuel costs, fuel expense, fuel supply, fuel commitments, related expenses, fossil fuel
5) Nuclear waste disposal costs: nuclear decommissioning, nuclear power, nuclear fuel, spent nuclear, nuclear plant, decommissioning costs, nuclear waste, waste policy, disposal of spent, nuclear business
6) Financial statement information: dollars in millions, ended december, income taxes, financial statements, accompanying notes, millions except, year ended, pension benefits, consolidated balance, consolidated financial
7) Restaurant business growth: company operated, operated restuarants, company owned, franchised restuarants, company opened, operated restuarants, restuarants at december, franchisees opened, owned and operated, opened restuarants
8) Derivatives and hedging activities: derivative financial, financial instruments, trading purposes, derivative instruments, instruments for trading, hold or issue, issue derivative, enter into, hedging activities, speculative purposes
9) Real estate loan operations: real estate, national bank, estate loans, estate investment, loan bank, home loan, federal home, investment trust, trust REIT, mortgage loans
10) Accounts receivable and doubtful accounts: accounts receivable, doubtful accounts, allowances for doubtful, valuation allowances, vendor allowances, product returns, allowances include, estimated allowances, uncollectible amounts, based on historical
11) Corporate spin-offs: prior to the spin, spin off from, spin off transaction, financial statements, since the spin, adjusted to reflect, completed the spin, after the spin, common stock, been adjusted
12) Segment performance: generation segment, discussed below, segment as discussed, because of items, segment because, merchant services, other operations, increased million, items detailed, maintenance expenses
13) Foreign country risks: republic of china, united states, located in china, agreement with, entered into, future inflation, inflation in china, conduct business, business in china, china may inhibit
14) Laser products: excimer laser, laser system, laser vision, laser technology, laser printers, laser based, laser beam, laser products, capital expenditures, discontinued operations
15) Digital technology and services: digital media, internet access, high speed, digital signal, analog and digital, internet services, services include, cable television, digital imaging, internet based
16) Long term assets: property and equipment, equipment property, property plant, stated at cost, intellectual property, equipment consisted, carried at cost, long lived, intangible assets, assets include
17) Franchise revenue recognition: franchise fees, franchise agreements, initial franchise, franchise royalties, franchise operations, franchise revenues, franchise rights, development fees, intangible assets, brand name
18) Business structure: holding company, bank holding, loan holding, company under, financial holding, holding corporation, utility holding, holding companies, unrealized holding, holding gains
19) Debt issuance: convertible debenture, common stock, subordinated debenture, agreement with, purchase agreement, principal amount, note or debenture, debenture holders, company issued, debenture offering
20) Media and entertainment: interactive entertainment, entertainment software, entertainment company, entertainment group, agreement dated, entertainment industry, entertainment services, company acquired, lease agreement, digital entertainment
21) Food products and services: food service, drug administration, food and drug, food products, food packaging, food and beverage, food processors, food distribution, quality food, food processing
22) Merger activities: merger with, merger agreement, plan of merger, prior to the merger, completed a merger, merger related, agreement and plan, proposed mergers, approved the merger, closing of the merger

Continued on next page

23) R&D partnerships: pharmaceutical products, collaboration agreement, agreement between, pharmaceutical company, collaboration with, entered into, pharmaceutical services, license agreement, research collaboration, between the company
24) Consolidated financial information: consolidated statements, cash flows, statements of cash, subsidiaries consolidated, corporation consolidated, consolidated balance, balance sheets, comprehensive income, share amounts, company consolidated
25) Advertising expenses: advertising costs, company expenses, advertising expense, costs are expensed, expense as incurred, first time, costs as incurred, online advertising, advertising revenue, advertising and promotion
26) Country risks: country to country, from country, united states, vary from, country basis, country by country, based on the country, outside the united, country risk, widely from
27) Joint venture agreements: joint venture, joint and several, full and unconditional, entered into, venture agreement, agreement with, venture between, venture partners, joint plan, formed a joint
28) Credit card operations: credit card, debit card, card transactions, card receivables, card services, card processing, gift card, card loans, interchange fees, card revenue,
29) Fair value/cash flow hedging: interest rate, rate swap, swap agreement, entered into, notional amount, fair value, floating rate, swap transactions, currency swap, hedge accounting
30) Purchase agreements: common units, each purchaser, authorized purchaser, units purchased, agreement with, with the purchaser, entered into, third party, affiliated purchaser, place an order
31) Legal proceedings: district court, southern district, district of new york, bankruptcy court, court of appeals, northern district, court has not yet ruled, eastern district, supreme court, appeals for the district
32) Strategic alliances: strategic alliance, alliance with, entered into, agreement with, alliance agreement, alliance partners, into a strategic, company entered, license agreement, ventures sold
33) Credit agreements: jpmorgan chase, kinder morgan, vice president, agreement with, entered into, credit facility, facility with, chase manhattan, credit agreement, company entered
34) Patent infringement and rights: patent applications, patent and trademark, trademark office, regulations provide, patent infringement, procedure for challenging, patent rights, control presumption, rebuttable control, filed a patent
35) Stock option plans: stock option, shall not exceed, board of directors, this agreement, shall be entitled, company shall, shall be entitled, option shall, meaning set forth, shall become
36) Share capital: preferred stock, redeemable preferred, mandatorily redeemable, cumulative redeemable, investing activities, convertible preferred, series A cumulative, common stock, series A preferred, capital securities
37) Partnership arrangements: general partner, limited partner, sole general, managing general, operating partnership, managing partner, limited partnerships, partner interest, executive officers, responsible for managing
38) Equity ownership and control: life insurance, common stock, consolidated statements, property trust, insurance company, agreement with, limited partnership, property limited, year ended, december ended
39) Share capital transactions: real estate, preferred stock, convertible preferred, estate loans, commercial real, series A convertible, series A preferred, preferred units, estate construction, construction loans
40) Segment performance: segment information, operating segment, business segment, segment consists, operating income, segment performance, performance based, reportable segment, segment reporting, segment results
41) Securitized/guaranteed securities: guaranteed securities, fannie mae and freddie, mortgage backed, backed securities, farmer mac guaranteed, mortgage loans, preferred stock, government sponsored, securities issued, freddie mac preferred
42) Environmental risks: mining operations, environmental liabilities, environmental remediation, environmental regulation, environmental risks, environmental laws, environmental compliance, environmental protection, environmental costs, mining claims
43) Foreign currency risks: functional currencies, foreign currencies, local currencies, asia pacific, foreign subsidiaries, denominated in foreign, company's foreign, foreign operations, consolidated statements, respective local
44) Geographic locations: located in, united stated, primarily in the united, latin america, united kingdom, canada and europe, located in the united, throughout the united, south america, asia pacific
45) Short-term credit facilities: credit facility, revolving credit, senior secured, entered into, secured credit, senior credit, term loan, credit agreement, secured revolving, term debt
46) End-of-year transactions: ending december, terminates on december, dated december, december we acquired, dated december, investment on december, year ending, company acquired, invested on december, payment was due on december

Continued on next page

47) Reference to quantitative information: table of contents, this table, significant table, based on table, guidance table, this guidance, include table, goodwill table, loans table, purchase table
48) Consolidated financial information: subsidiaries consolidated, consolidated statements, balance sheets, consolidated balance, comprehensive income, statements of income, subsidiaries are listed, company and subsidiaries, statements of comprehensive, ended december
49) Corporate spin-offs: prior to the spin, following the spin, completed the spin, investment corporation, spin off from, with the spin, result of the spin, plan to spin, principal holdings, entered into
50) Aircraft leasing agreements: commercial aircraft, aircraft engines, boeing aircraft, entered into, aircraft maintenance, operating lease, lease agreement, additional aircraft, aircraft manufacturers, credit corporation
51) Gaming operations: gaming license, gaming operations, mississippi gaming, gaming machines, gaming commission, casino gaming, gaming control, native american, indian gaming, gaming taxes
52) Mining operations: ounces of gold, gold bank, gold and silver, gold mining, copper gold, mining claims, million ounces, square feet, into gold, gold production
53) Cable/television operations: cable television, television stations, television systems, television industry, cable systems, television programming, cable programming, fiber optic, cable operators, cable act imposed
54) Aircraft manufacturing operations: mcdonnell douglas, boeing aircraft, boeing company, commercial aircraft, boeing mcdonell, lockheed martin, agreement between, customers include, savings bank, sales to boeing
55) Patient and nursing services: skilled nursing, nursing facilities, assisted living, nursing homes, nursing home, physical occupational, nursing care, living facilities, home health, real estate
56) Marine operations: insurance carriers, marine transportation, marine services, offshore marine, other carriers, licensed insurance, marine containers, maintain insurance, self insure, marine insurance
57) Agricultural operations: crop insurance, crop hail, crop production, agricultural partnerships, crop nutrient, insurance business, crop yields, named peril, crop drying, agricultural market
58) Hotel and lodging operations: interstate hotels, hotels resorts, service hotels, full service, united states, hotels and resorts, hotel properties, managed hotel, hotels are located, ownership of management
59) Floral products: crop insurance, crop protection, fresh cut flowers, floral products, specialty retailers, crop nutrient, floral services, named peril, crop year, brooding and weed
60) Gaming regulations and violations: nevada gaming, vegas nevada, gaming authorities, nevada commission, authorities at any time, current stock, examined by the nevada, district court, court for the district, nevada board
61) Measurement of natural gas properties: natural gas properties, natural gas reserves, inherently precise, reserves are inherently, business consists, cost method, full cost, method of accounting, involves a high, estimates of oil and natural
62) Apparel manufacturing: women's apparel, outlet stores, czech republic, apparel group, apparel manufacturers, weekly basis, united states, business segments, dominican republic, mens apparel
63) Utility operations: commodity price, square miles, electric service, service to communities, furnishes electric, communities in square, plans in early, companies expect, price risk, miles of western
64) Gold mining operations: gold project, gold and silver, gold prices, ounces of gold, gold mine, entered into, gold mineralization, northern territy, gold exploration, gold production

This table presents the 10 representative bigrams for each combined topic. The topic number and label are presented in bold and are based on the top 20 bigrams from a set of 334 representative sentences for each topic. The bigrams presented are the 10 most common n-grams after excluding redundant phrases (e.g., "compared with" and "compared to" in topic 2). For ease of exposition, industry-specific topics are placed at the end (topics 50 through 64).

Appendix D Formulation of the Var-Gamma Test

The Var-Gamma test is a test of the difference of X^2 distributed variables that are independent and random. This test is well-suited for our analyses as the test statistics from Fisher's method follow a X^2_{2k} distribution, where k is the number of observations. Let X_1 and X_2 be independent and distributed following X^2_{2k}. The moment generating function of X^2_{2k} is $(1-2t)^{-k}$. Thus, the moment generating function of $X_1 - X_2$ is derived as:

$$\begin{aligned} M_{X_1}(t)M_{X_2}(t) &= (1-2t)^{-k}(1+2t)^{-k}, \\ &= ((1-2t)(1+2t))^{-k}, \\ &= (1-4t^2)^{-k}, \\ &= \left(\frac{1}{1-4t^2}\right)^k, \\ &= \left(\frac{\frac{1}{4}}{\frac{1}{4}-t^2}\right)^k. \end{aligned} \tag{4}$$

Next, we note that the moment generating function of the Variance Gamma distribution is given by $e^{\mu t}\left(\frac{\alpha^2-\beta^2}{\alpha^2-(\beta+t)^2}\right)^{\lambda}$. Thus, the function in equation (4) is a special case of the moment generating function of the variance gamma distribution with $\mu = 0$, $\beta = 0$, $\lambda = k$, and $\alpha = \frac{1}{2}$. The pdf of the resulting distribution is therefore

$$f(z) = \frac{1}{2^k\sqrt{\pi}\,\Gamma(k)}|z|^{k-\frac{1}{2}}K_{k-\frac{1}{2}}(|z|),$$

where $\Gamma(k)$ is the gamma function and $K_{k-\frac{1}{2}}$ is the modified Bessel function of the second kind. We can then conduct our statistical test as follows:

$$\mathbb{P}\left(X_1 < X_2\right) = \mathbb{P}\left(X_1 - X_2 < 0\right) = \int_{-\infty}^{X_1-X_2} f(z) = \frac{1}{2^k\sqrt{\pi}\,\Gamma(k)}|z|^{k-\frac{1}{2}}K_{k-\frac{1}{2}}(|z|)dz. \tag{5}$$

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): 217–250.

BAO, Y. and A. DATTA. 'Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures'. Management Science 60 (2014): 1371–1391.

BAO, Y., B. KE, B. LI, Y. YU, and J. ZHANG. 'Detecting Accounting Frauds in Publicly Traded U.S. Firms: New Perspectives and New Methods'. Working paper. 2018.

BAUGUESS, S. W. 'Use of AI and Machine Learning in Market Risk Assessment'. Remarks at the Practicing Law Institute 2018 SEC Speaks Conference. Feb. 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): 443–465.

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): 441–454.

BELLSTAM, G., S. BHAGAT, and J. A. COOKSON. 'A Text-Based Analysis of Corporate Innovation'. Working paper. 2017.

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.

— '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): 248–252.

BONSALL IV, 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, Ohio State University. 2014.

BOUKUS, E. and J. V. ROSENBERG. 'The Information Content of FOMC Minutes'. Working paper, Federal Reserce 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 (2017). Forthcoming.

BRAZEL, J. F., K. L. JONES, and M. F. ZIMBELMAN. 'Using Nonfinancial Measures to Assess Fraud Risk'. Journal of Accounting Research 47 (2009): 1135–1166.

BROWN, S. V. and J. W. TUCKER. 'Large-Sample Evidence on Firms Year-over-Year MD&A Modifications'. Journal of Accounting Research 49 (2011): 309–346.

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.

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): 164–175.

— 'Detecting Management Fraud in Public Companies'. Management Science 56 (2010b): 1146–1160.

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. 2009: 288–296.

COSTER, W. and D. KAUCHAK. 'Simple English Wikipedia: A New Text Simplification Tool'. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Shortpapers (2011): 665–669.

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): 11600–11605.

DECHOW, P. M., W. GE, C. R. LARSON, and R. G. SLOAN. 'Predicting Material Accounting Misstatement in Accounting'. Contemporary Accounting Research 28 (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 Action 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): 682–696.

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): 221–245.

EAGLESHAM, J. 'Accounting Fraud Targeted'. Wall Street Journal (May 2013).

EICKHOFF, M. and N. NEUSS. 'Topic Modelling Methodology: Its Use in Information Systems and Other Managerial Disciplines'. the 25th European Conference on Information Systems (ECIS). 2017: 1327–1347.

FARBER, D. B. 'Restoring Trust after Fraud: Does Corporate Governance Matter?' The Accounting Review 80 (2005): 539–561.

FARRELL, A. M., J. H. GRENIER, and J. LEIBY. 'Scoundrels or Stars? Theory and Evidence on the Quality of Workers in Online Labor Markets'. The Accounting Review 92 (2017): 93–114.

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): 107–142.

FILES, R. 'SEC Enforcement: Does Forthright Disclosure and Cooperation Really Matter?' Journal of Accounting and Economics 53 (2012): 353–374.

FISHER, R. A. Statistical Methods for Research Workers. 4th ed. Edinburgh: Oliver & Boyd, 1932.

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

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): 234–269.

GUTIERREZ, E. F., J. KRUPA, M. MINUTTI-MEZA, and M. VULCHEVA. 'How Useful Are Auditors' Going Concern Opinions as Predictors of Default?' 2017.

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): 1487–1519.

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): 349–392.

HOFFMAN, M., F. R. BACH, and D. M. BLEI. 'Online Learning for Latent Dirichlet Allocation'. Advances in Neural Information Processing Systems. 2010: 856–864.

HUANG, A., R. LEHAVY, A. ZANG, and R. ZHENG. 'Analyst Information Discovery and Information Interpretation Roles: A Topic Modeling Approach'. Management Science (2017). Forthcoming.

HUMPHERYS, S. L., K. C. MOFFIT, M. B. BURNS, J. K. BURGOON, and W. F. FELIX. 'Identification of Fraudulent Financial Statements Using Linguistic Credibility Analysis'. Decision Support Systems 50 (2011): 585–594.

JANES, H., G. LONGTON, and M. PEPE. 'Accommodating Covariates in ROC Analysis'. The Stata Journal 9 (2009).

KIM, I. and D. J. SKINNER. 'Measuring Securities Litigation Risk'. Journal of Accounting and Economics 53 (2012): 290–310.

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'. The 26th XBRL International Conference. Dublin, Ireland, Apr. 2013. URL: https://www.youtube.com/watch?feature=player_detailpage&v=EdfEEmXcYXU.

LI, F. 'The Information Content of Forward-Looking Statements in Corporate Filings - A Naïve Bayesian Machine Learning Approach'. Journal of Accounting Research 48 (2010a): 1049–1102.

— 'Textual Analysis of Corporate Disclosures: A Survey of the Literature'. Journal of Accounting Literature 29 (2010b): 143–165.

— 'Annual Report Readability, Current Earnings, and Earnings Persistence'. Journal of Accounting and Economics 45 (2008).

LI, H. 'Repetitive Disclosures in the MD&A'. Dissertation, University of Toronto. 2014.
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.
— 'Textual Analysis in Accounting and Finance: A Survey'. Journal of Accounting Research 54 (2016): 1187–1230.
MCAULIFFE, J. D. and D. M. BLEI. 'Supervised Topic Models'. Advances in Neural Information Processing Systems. 2008: 121–128.
MCLEAN, B. 'Is Enron Overpriced?' Fortune Magazine (Mar. 2001).
MIKOLOV, T., K. CHEN, G. CORRADO, and J. DEAN. 'Efficient Estimation of Word Representations in Vector Space'. Workshop at ICLR. 2013.
MURPHY, M. and K. TYSIAC. 'Data Analytics Helps Auditors Gain Deep Insight'. Journal of Accountancy (Apr. 2015): 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): 665–675.
PENNINGTON, J., R. SOCHER, and C. D. MANNING. 'GloVe: Global Vectors for Word Representation'. Empirical Methods in Natural Language Processing. 2014: 1532–1543.
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): 221–245.
PURDA, L. and D. SKILLICORN. 'Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection'. Contemporary Accounting Research 32 (2015): 1193–1223.
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): 209–228.
RENNEKAMP, K. 'Processing Fluency and Investor's Reactions to Disclosure Readability'. Journal of Accounting Research 50 (2012): 1319–1354.
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): 437–485.
ROGERS, J. L., A. V. BUSKIRK, and S. L. ZECHMAN. 'Disclosure Tone and Shareholder Litigation'. The Accounting Review 86 (2011): 2155–2183.
SINGER, Z. and J. ZHANG. 'Auditor Tenure and the Timeliness of Misstatement Discovery'. The Accounting Review (2018). Forthcoming.
WANG, Y., L. WANG, Y. LI, D. HE, and T.-Y. LIU. 'A Theoretical Analysis of NDCG Type Ranking Measures'. Proceedings of the 26th Annual Conference on Learning Theory. Vol. 30. Proceedings of Machine Learning Research. Princeton, NJ, USA, 2013: 25–54.

Figure 1: Combined Topic Distribution and Irregularity Restatement Prediction

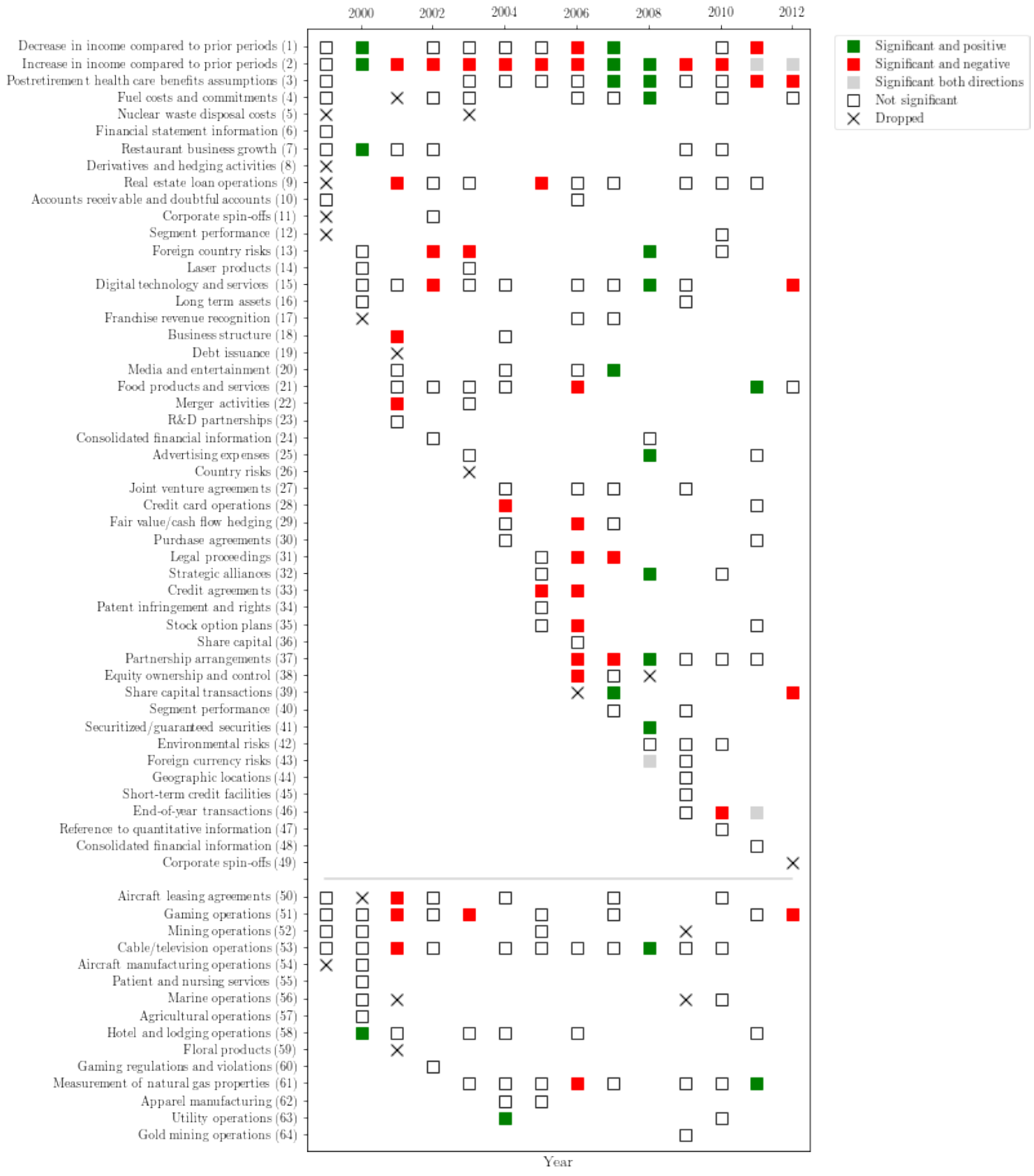

Figure 1: Combined Topic Distribution and Irregularity Restatement Prediction (Continued)

This chart depicts the calendar years in which a combined topic is present in our corpus of 10-K filings. We also illustrate the predictive ability of each topic in detecting misstatements involving irregularity restatements. We assess the predictive ability by estimating yearly logit regressions of our *misreporting* indicator variable on a vector of disaggregated subtopics (i.e., those topics that are associated with a given combined topic) present within each year. We orthogonalize all subtopics to 2-digit SIC industries to control for industry effects. Topics that are present in a given year are denoted using a square box; topics that are present but dropped from our prediction model due to collinearity are denoted as an X. We color code the square boxes based on the direction and statistical significance of the disaggregated subtopics from our logit regressions. The square boxes are color coded green (red) if the subtopics positively (negatively) predict misstatements involving irregularity restatements. The box is color coded grey if the subtopics are statistically significant but with ambiguous direction (multiple subtopics loading in opposing direction). White square boxes indicate that the disaggregated subtopics are insignificant in that particular year. Topics below the line (topic 50 to topic 64) are industry-specific topics.

Figure 2: Combined Topic Distribution and AAER Prediction

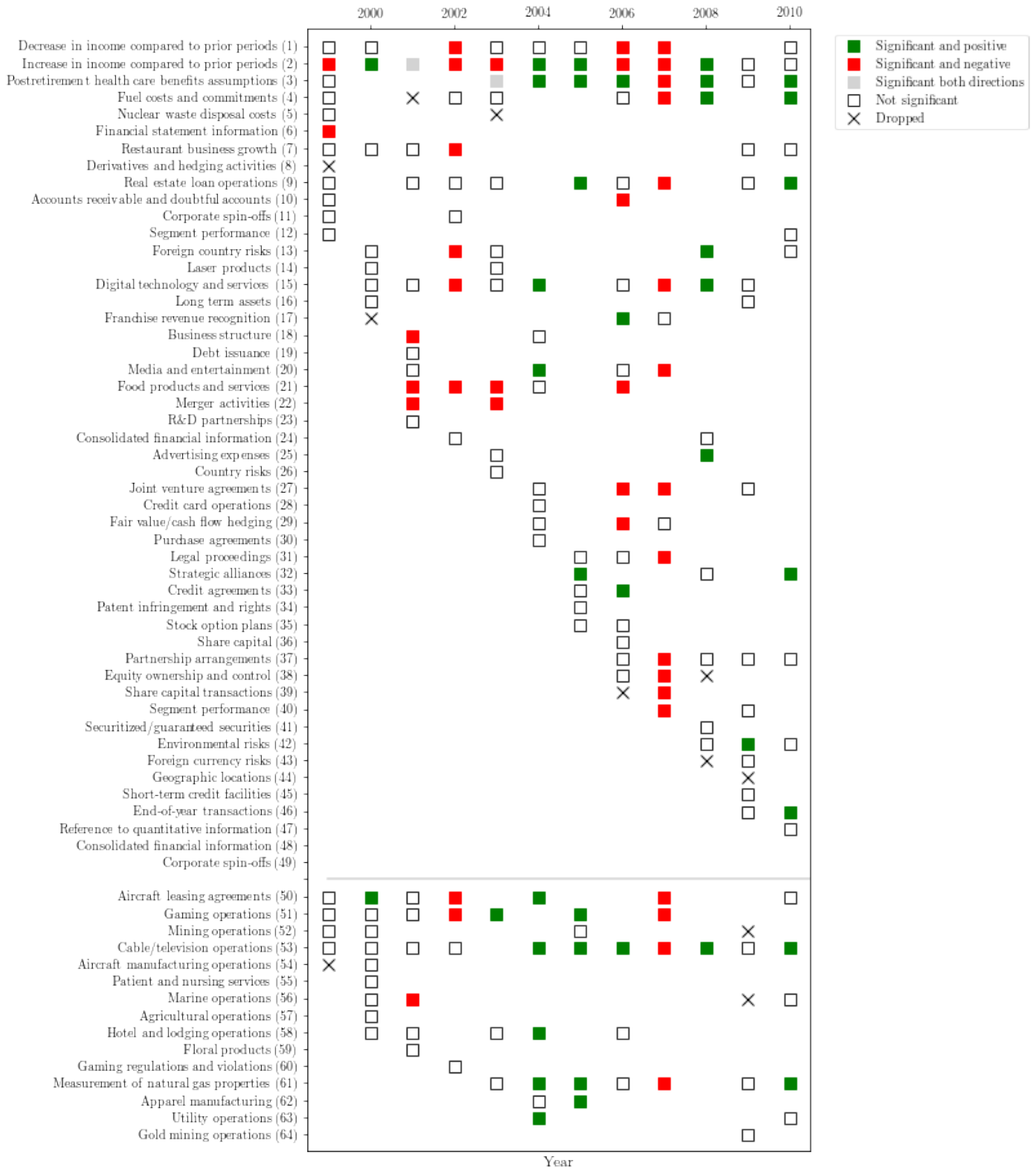

Figure 2: Combined Topic Distribution and AAER Prediction (Continued)

This chart depicts the yearly presence of each combined topic and the predictive ability of each topic in detecting misstatements involving SEC enforcement actions (AAERs). Topics that are present in a given year are denoted using a square box; topics that are dropped from our prediction model due to collinearity are denoted using an X. The square boxes are color coded based on the direction and statistical significance of the 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 2-digit SIC industries to control for industry effects. The square boxes are color coded green (red) if the subtopics positively (negatively) predict misstatements involving irregularity restatements. The box is color coded grey if the subtopics are statistically significant but with ambiguous direction (multiple subtopics loading in opposing directions). White square boxes indicate that the disaggregated subtopics are insignificant in the respective year. Topics below the line (topic 50 to topic 64) are industry-specific topics.

Table 1: Distribution of AAERs and Irregularity Restatements by Year

	AAER			Irregularity Restatement		
Year	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	16	0.98	1,634	17	1.04
1997	2,250	20	0.89	2,250	16	0.71
1998	2,308	37	1.60	2,308	12	0.52
1999	2,195	46	2.10	2,195	13	0.59
2000	2,041	50	2.45	2,041	21	1.03
2001	2,021	43	2.13	2,021	18	0.89
2002	2,391	50	2.09	2,391	27	1.13
2003	2,936	57	1.94	2,936	53	1.81
2004	2,843	49	1.72	2,843	70	2.46
2005	2,678	38	1.42	2,678	65	2.43
2006	2,608	18	0.69	2,608	78	2.99
2007	2,549	15	0.59	2,549	53	2.08
2008	2,535	8	0.32	2,535	48	1.89
2009	2,564	5	0.20	2,564	65	2.54
2010	2,424	1	0.04	2,424	48	1.98
2011				2,330	44	1.89
2012				2,178	45	2.07
Total	37,806	459	1.21	42,314	697	1.65

This table reports the distribution of our sample observations as well as the distribution of AAERs and irregularity restatements by year. Our AAER sample excludes observations from the 2011 and 2012 years since there are no AAER events during those two years. The first (fourth) column reports the total number of firm observations per year for the AAER (irregularity restatement) sample. The second (fifth) column reports the frequency of AAERs (irregularity restatements) per year, while the third (sixth) column reports the percent of observations that have an AAER (irregularity restatement) per year. The final row reports overall frequencies and percentages for the full sample.

Table 2: Summary Statistics of Financial Variables for Misstated versus Non-Misstated Firm-Years

	No AAER		AAER			No Irreg. Restate.		Irreg. Restate.		
Variable	Mean	Std. Dev.	Mean	Std. Dev.	Difference	Mean	Std. Dev.	Mean	Std. Dev.	Difference
$\log(TotalAssets)$	5.68	1.95	6.55	1.82	0.875^{***}	5.76	1.97	5.74	2.02	-0.0262
$RSSTAccruals$	0.0150	0.296	0.0305	0.234	0.0155	0.0153	0.285	0.0081	0.273	-0.00719
$\Delta Receivables$	0.0093	0.0697	0.0206	0.0684	0.0114^{***}	0.0091	0.0677	0.0125	0.0660	0.00339
$\Delta Inventory$	0.0053	0.0537	0.0138	0.0615	0.00853^{***}	0.0058	0.0523	0.0029	0.0570	-0.00281
$\%SoftAssets$	0.534	0.241	0.656	0.200	0.122^{***}	0.535	0.241	0.549	0.252	0.0143
$\Delta CashSales$	0.267	12.9	0.202	0.398	-0.0655	0.212	14.4	0.0599	3.25	-0.152
$\Delta ReturnOnAssets$	-0.0030	0.317	-0.0247	0.194	-0.0218^{**}	-0.0039	0.304	-0.0127	0.264	-0.00877
$ActualIssuance$	0.925	0.263	0.965	0.184	0.0400^{***}	0.923	0.266	0.941	0.235	0.0181^{**}
$OperatingLeases$	0.866	0.341	0.891	0.312	0.0254^{*}	0.868	0.339	0.882	0.322	0.0146
$Book\text{-}To\text{-}Market$	0.501	6.69	0.537	0.672	0.0360	0.506	6.38	0.507	1.20	0.00095
$Lag(Mkt\text{-}AdjReturn)$	0.106	0.982	0.195	0.984	0.0884^{*}	0.102	0.945	0.0667	0.876	-0.0351
$Merger$	0.191	0.393	0.349	0.477	0.157^{***}	0.193	0.395	0.212	0.409	0.0189
$BigNAuditor$	0.824	0.381	0.889	0.315	0.0646^{***}	0.816	0.387	0.740	0.439	-0.0759^{***}
$Mid\text{-}sizeauditor$	0.0858	0.280	0.0610	0.240	-0.0248^{**}	0.0896	0.286	0.115	0.319	0.0252^{**}
$TotFinancing$	0.0431	0.237	0.0496	0.176	0.00649	0.0408	0.233	0.0908	0.330	0.0500^{***}
$ExanteFinancing$	-0.0731	1.65	0.0298	0.248	0.103^{***}	-0.0613	1.57	-0.231	1.26	-0.170^{***}
$Restructuring$	0.205	0.404	0.294	0.456	0.0888^{***}	0.217	0.412	0.323	0.468	0.106^{***}

This table reports summary statistics of our financial variables for misstated and non-misstated firm-years for the sample of AAERs and irregularity restatements. We conduct two-tailed t-tests of the differences in means for the misstated and non-misstated firm-years in each sample. Appendix B provides the definitions of the financial variables, while Section 3.1.1 discusses the data sources for the AAER and irregularity restatement samples. The irregularity restatement (AAER) sample consists of $42,314$ ($37,806$) firm-years from January 1st, 1994 through December 31st, 2012 (December 31st, 2010), of which 697 (459) involve a material accounting misstatement. The *Restructuring* variable is valid only for the post-1999 period since restructuring charges were not separately reported in Compustat prior to 2000. The restatement (AAER) sample for the post-1999 period consists of $32,098$ ($27,590$) firm-years of which 635 (334) involve a material accounting misstatement. 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: Fisher Tests & Pooled ROC AUC Statistics (AAER Sample)

Specification	Fisher Statistic	p-value	Pooled ROC AUC
F-score	151.94^{***}	< 0.001	0.742^{***}
topic	113.90^{***}	< 0.001	0.700^{***}
topic and *F-score*	194.35^{***}	< 0.001	0.769^{***}

Panel B: Difference Tests (AAER Sample)

	Measure	*F-score*	*topic* and *F-score*
topic	Var-Gamma	-38.04^{***} (< 0.001)	-80.45^{***} (< 0.001)
	ROC AUC	-0.042^{*} (0.057)	-0.069^{***} (< 0.001)
topic and *F-score*	Var-Gamma	42.41^{***} (< 0.001)	
	ROC AUC	0.027^{***} (0.002)	

Panel C: Fisher Tests & Pooled ROC AUC Statistics (Irregularity Restatement Sample)

Specification	Fisher Statistic	p-value	Pooled ROC AUC
F-score	35.19	0.165	0.589^{***}
topic	95.53^{***}	< 0.001	0.616^{***}
topic and *F-score*	85.84^{***}	< 0.001	0.630^{***}

Panel D: Difference Tests (Irregularity Restatement Sample)

	Measure	*F-score*	*topic* and *F-score*
topic	Var-Gamma	60.35^{***} (< 0.001)	9.69 (0.355)
	ROC AUC	0.027^{*} (0.065)	-0.014^{*} (0.075)
topic and *F-score*	Var-Gamma	50.66^{***} (< 0.001)	
	ROC AUC	0.041^{***} (< 0.001)	

This table provides comparative out-of-sample tests of our prediction models based on *topic* and financial metrics (denoted as *F-score*). Section 3.2.3 describes the measurement of *topic*, while Appendix B defines the financial variables. Panels A and C present test statistics using Fisher's [1932] method, associated p-values, and pooled ROC AUC values (with bias corrected p-values based on bootstrap clustered by year) for the AAER and irregularity restatement samples, respectively (see Section 3.1.1 for a description of our data sources). The test statistics are based on an aggregation of p-values or pooled out-of-sample predictions generated by regressions of *misreport* on $p_{misreport}$ generated by the rolling five-year windows. Degrees of freedom for the Fisher's method tests in panels A and C are 24 and 28, respectively. Panels B and D present the Var-Gamma tests of the differences in the Fisher statistics across models (see Appendix D) as well as non-parametric Wald tests of the differences in the AUCs for the AAER and irregularity restatement samples, respectively. The Var-Gamma tests are based on 12 and 14 prediction years, respectively, for the AAER and irregularity restatement samples, while the Wald tests are based on 1,000 bootstrap iterations with clustering by year. The panels report 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 4: Summary Statistics of Textual Style Features for Misstated versus Non-Misstated Firm-Years

	No AAER		AAER			No Irreg. Restate		Irreg. Restate.		
Variable	Mean	Std. Dev.	Mean	Std. Dev.	Difference	Mean	Std. Dev.	Mean	Std. Dev.	Difference
Length										
ParsedSize	171707	101909	165765	87667	−5942	177540	104221	237193	115888	59653***
SentenceLength	23.8	2.29	23.5	2.24	−0.302***	23.9	2.31	24.6	2.21	0.626***
Complexity										
WordStddev	3.07	0.0707	3.08	0.0844	0.00773*	3.07	0.0699	3.07	0.0626	0.00319
ParagraphStddev	4.06	3.78	3.82	2.93	−0.234*	5.12	11.4	6.92	15.8	1.80***
Variation										
Repetitions	0.0794	0.0488	0.0810	0.0492	0.00159	0.0791	0.0481	0.0994	0.0545	0.0203***
SentenceStddev	16.5	4.66	16.6	4.18	0.0290	16.4	4.49	16.3	4.21	−0.117
TypeTokenRatio	0.126	0.0540	0.127	0.0608	0.00099	0.124	0.0549	0.101	0.0378	−0.0237***
Readability										
Coleman-LiauIndex	14.5	0.691	14.6	0.722	0.0880***	14.5	0.689	14.4	0.564	−0.0662***
Fog	17.8	1.44	17.6	1.45	−0.131*	17.9	1.47	18.4	1.23	0.501***
Tense										
%ActiveVoice	0.606	0.0681	0.598	0.0709	−0.00808**	0.611	0.0680	0.625	0.0596	0.0142***
%PassiveVoice	0.0314	0.0208	0.0309	0.0157	−0.00046	0.0313	0.0205	0.0312	0.0174	−0.00003
Word Choice										
%Negative	0.0126	0.0050	0.0127	0.0053	0.00006	0.0129	0.0050	0.0161	0.0048	0.00324***
%Positive	0.0064	0.0021	0.0061	0.0016	−0.00028***	0.0064	0.0021	0.0065	0.0017	0.00008
Emphasis										
AllCaps	567	515	579	472	12.4	556	506	702	688	146***
ExclamationPoints	0.371	7.07	0.0566	0.513	−0.315***	0.354	6.72	0.475	3.19	0.121
QuestionMarks	0.0930	2.36	0.0719	0.566	−0.0211	0.0890	2.30	0.347	4.78	0.258
Processing										
log (*Bullets*)	5.22	2.30	5.06	2.17	−0.163	5.19	2.31	5.47	2.13	0.280***
Header	1452	462	1441	161	−11.1	1429	452	1475	252	45.9***
Newlines	1591	1694	1201	1057	−390***	1698	1795	2144	2014	446***
Tags	472197	712325	290716	462449	−181481***	555212	795106	793039	874145	237827***

This table presents summary statistics of our textual style variables for misstated and non-misstated firm-years for the AAER and irregularity restatement samples. We conduct two-tailed t-tests of the difference in means for misstated and non-misstated firm-years in each sample. Appendix B provides the definitions of the textual style variables, while Section 3.1.1 discusses the data sources for the AAER and irregularity restatement samples. The irregularity restatement (AAER) sample consists of 42,314 (37,806) firm-years from January 1st, 1994 through December 31st, 2012 (December 31st, 2010), of which 697 (459) involve material accounting misstatement. 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 5: Out-of-Sample Prediction Analysis for *topic* and *Style*

Panel A: Fisher Tests & Pooled ROC AUC Statistics (AAER Sample)

Specification	Fisher Statistic	p-value	Pooled ROC AUC
Style	34.44^{*}	0.077	0.684^{***}
topic	113.90^{***}	< 0.001	0.700^{***}
topic and *Style*	93.44^{***}	< 0.001	0.735^{***}

Panel B: Difference Tests (AAER Sample)

	Measure	*Style*	*topic* and *Style*
topic	Var-Gamma	79.47^{***} (< 0.001)	20.47^{**} (0.038)
	ROC AUC	0.016 (0.180)	-0.035^{***} (0.001)
topic and *Style*	Var-Gamma	59.00^{***} (< 0.001)	
	ROC AUC	0.051^{***} (< 0.001)	

Panel C: Fisher Tests & Pooled ROC AUC Statistics (Irregularity Restatement Sample)

Specification	Fisher Statistic	p-value	Pooled ROC AUC
Style	128.42^{***}	< 0.001	0.663^{***}
topic	95.53^{***}	< 0.001	0.616^{***}
topic and *Style*	163.37^{***}	< 0.001	0.669^{***}

Panel D: Difference Tests (Irregularity Restatement Sample)

	Measure	*Style*	*topic* and *Style*
topic	Var-Gamma	-32.89^{***} (0.002)	-67.84^{***} (< 0.001)
	ROC AUC	-0.047^{***} (< 0.001)	-0.054^{***} (< 0.001)
topic and *Style*	Var-Gamma	34.94^{***} (0.001)	
	ROC AUC	0.007 (0.284)	

In this table, we report comparative test results of the predictive power of models based on *topic* and textual style features (denoted as *Style*). Section 3.2.3 describes the measurement of *topic*, while Appendix B defines the textual style variables. Panels A and C present test statistics using Fisher's [1932] method, associated p-values, and pooled ROC AUC values (with bias corrected p-values based on bootstrap clustered by year) for the AAER and irregularity restatement samples, respectively (see Section 3.1.1 for a description of our data sources). The test statistics are based on an aggregation of p-values or pooled out-of-sample predictions generated by regressions of *misreport* on $p_{misreport}$ generated by the rolling five-year windows. Degrees of freedom for the Fisher's method tests in panels A and C are 24 and 28, respectively. Panels B and D present the Var-Gamma tests of the differences in the Fisher statistics across models (see Appendix D) as well as non-parametric Wald tests of the differences in the AUCs for the AAER and irregularity restatement samples, respectively. The Var-Gamma tests are based on 12 and 14 prediction years, respectively, for the AAER and irregularity restatement samples, while the Wald tests are based on 1,000 bootstrap iterations with clustering by year. The panels report 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 for *topic*, *F-score*, and *Style*

Panel A: Fisher Tests & Pooled ROC AUC Statistics (AAER Sample)

Specification	Fisher Statistic	p-value	Pooled ROC AUC
F-score and *Style*	164.44^{***}	< 0.001	0.757^{***}
topic, *F-score*, and *Style*	202.62^{***}	< 0.001	0.778^{***}
topic	113.90^{***}	< 0.001	0.700^{***}
topic and *F-Score*	194.35^{***}	< 0.001	0.769^{***}

Panel B: Difference Tests (AAER Sample)

	Measure	*F-score* and *Style*	*topic*, *F-score*, and *Style*
topic	Var-Gamma	-50.54^{***} (< 0.001)	-88.71^{***} (< 0.001)
	ROC AUC	-0.057^{***} (0.004)	-0.078^{***} (< 0.001)
topic and *F-Score*	Var-Gamma	29.91^{***} (0.003)	-8.27 (0.393)
	ROC AUC	0.012 (0.193)	-0.008 (0.331)
topic, *F-score*, and *Style*	Var-Gamma	38.17^{***} (< 0.001)	
	ROC AUC	0.020^{**} (0.037)	

Panel C: Fisher Tests & Pooled ROC AUC Statistics (Irregularity Restatement Sample)

Specification	Fisher Statistic	p-value	Pooled ROC AUC
F-score and *Style*	141.46^{***}	< 0.001	0.667^{***}
topic, *F-score*, and *style*	167.62^{***}	< 0.001	0.670^{***}
topic	95.53^{***}	< 0.001	0.616^{***}
topic and *Style*	163.37^{***}	< 0.001	0.669^{***}

Panel D: Difference Tests (Irregularity Restatement Sample)

	Measure	*F-score* and *Style*	*topic*, *F-score*, and *Style*
topic	Var-Gamma	-45.92^{***} (< 0.001)	-72.08^{***} (< 0.001)
	ROC AUC	-0.051^{***} (< 0.001)	-0.054^{***} (< 0.001)
topic and *Style*	Var-Gamma	21.91^{**} (0.040)	-4.25 (0.684)
	ROC AUC	0.003 (0.570)	-0.001 (0.744)
topic, *F-score*, and *Style*	Var-Gamma	26.16^{**} (0.015)	
	ROC AUC	0.004 (0.417)	

Table 6: Out-of-Sample Prediction Analysis for *topic*, *F-score*, and *Style*

This table reports comparative tests of the out-of-sample prediction power of models based on *topic*, financial metrics (*F-score*), and textual style features (*Style*). Section 3.2.3 details the construction of *topic*, while Appendix B defines the financial and textual style variables. Panels A and C present test statistics using Fisher's [1932] method, associated p-values, and pooled ROC AUC values (with bias corrected p-values based on bootstrap clustered by year) for the AAER and irregularity restatement samples, respectively (see Section 3.1.1 for a description of our data sources). The test statistics are based on an aggregation of p-values or pooled out-of-sample predictions generated by regressions of *misreport* on $p_{misreport}$ generated by the rolling five-year windows. Degrees of freedom for the Fisher's method tests in panels A and C are 24 and 28, respectively. Panels B and D present the Var-Gamma tests of the differences in the Fisher statistics across models (see Appendix D) as well as non-parametric Wald tests of the differences in the AUCs for the AAER and irregularity restatement samples, respectively. The Var-Gamma tests are based on 12 and 14 prediction years, respectively, for the AAER and irregularity restatement samples, while the Wald tests are based on 1,000 bootstrap iterations with clustering by year. The panels report 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 7: Classification Performance of *topic* for AAERs and Irregularity Restatements

Panel A: Classification of AAERs

	Classification %			NDCG@k
	50th	90th	95th	99th
topic	72.54	18.60	11.25	0.097
F-score	71.16	23.86	14.04	0.141
Style	60.21	11.95	6.50	0.085
topic and *F-score*	74.07	32.07	17.24	0.192
topic and *Style*	74.47	19.40	11.27	0.123
F-score and *Style*	73.98	23.73	14.66	0.168
topic, *F-score*, and *Style*	75.09	31.50	21.44	0.176

Panel B: Classification of Irregularity Restatements

	Classification %			NDCG@k
	50th	90th	95th	99th
topic	65.19	19.74	10.49	0.107
F-score	58.33	15.96	9.95	0.135
Style	69.68	25.54	14.06	0.092
topic and *F-score*	63.31	21.40	11.30	0.107
topic and *Style*	70.61	24.55	15.56	0.135
F-score and *Style*	69.60	26.72	15.94	0.143
topic, *F-score*, and *Style*	70.18	28.05	16.83	0.163

This table reports the classification accuracy of our detection models using out-of-sample prediction scores. Panel A (Panel B) reports the results for the AAER (irregularity restatement) sample. For each prediction model, we report in the first three columns the average annual percentage of misstated 10-K filings that are accurately classified as misstated at the respective cut-offs at the 50th, 90th, and 95th percentiles of the predicted probability scores. The fourth column 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.