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

September 2016

*We thank Andrew Bauer, Amanda Convery, Paul Demeré, Shawn Gordon, Jing He, Shiva Rajgopal, Kristina Rennekamp, Gang Wang, and workshop participants at Baruch - CUNY, Columbia University, University of Illinois, 2015 AAA FARS Mid-year Meeting, 2015 AAA Annual Meeting, 2015 Conference on Convergence of Financial and Managerial Accounting Research, and the 2016 Conference on Investor Protection, Corporate Governance, and Fraud Prevention for their helpful comments. We also thank Xiao Yu for insightful comments on methodology and coding, and Stephanie Grant, Jill Santore, and Jingpeng Zhu for excellent research assistance.

Electronic copy available at: http://ssrn.com/abstract=2803733

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 models based on *topic* outperform models of commonly-used financial and textual style variables. Furthermore, we find that *topic* significantly improves the detection of high risk accounting misstatements when added to models based on financial and textual style metrics. These results are robust to alternative *topic* definitions and regression specifications.

Keywords: Topic, Disclosure, Latent Dirichlet Allocation, Financial Misreporting

Electronic copy available at: http://ssrn.com/abstract=2803733

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]). 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 style of management disclosures, finding that these features serve as useful warning signs of misreporting (e.g., Loughran and McDonald [2011], Hobson, Mayew, and Venkatachalam [2012], Larcker and Zakolyukina [2012], Purda and Skillicorn [2015]).[3]

Despite the usefulness of communication style in predicting accounting misstatements, the literature debates whether style features adequately capture managers' deliberate attempts to obfuscate information or engage in deceitful behavior (Bloomfield [2008], Bushee, Gow, and Taylor [2015]). Furthermore, as Loughran and McDonald [2016] highlight, commonly-used style 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 financial statement disclosures. This approach departs from prior text-based research

[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. [2010], Goel et al. [2010], Purda and Skillicorn [2015]), and measures of readability and textual complexity (Humpherys et al. [2011], Goel and Gangolly [2012]).

Electronic copy available at: http://ssrn.com/abstract=2803733

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. In addition, we investigate the incremental predictive power of *topic* in detecting accounting misstatements relative to the financial statement and textual style measures used in prior work.

Our focus on the thematic content of financial statements draws from prior research on linguistics, psychology, and deception, which suggests that communication topics are intentionally chosen and actively monitored during deception, whereas style choices are often subconscious or without explicit intent (McCornack [1992], Buller and Burgoon [1996], Chung and Pennebaker [2007]). This literature also finds that, even in the case of purposeful style choices, language features are difficult to classify as deceptive when individuals possess strategic goals to communicate misleading information (Douglas and Sutton [2003]). In this sense, the content of the message is likely to be a better predictor of deception than how the message is fashioned. We therefore conjecture that the topic content of financial statement disclosures will be incrementally informative in assessing the likelihood of misreporting, beyond language style features. We also expect *topic* to outperform the detection ability of quantitative accounting and financial 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., 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 search engines such as Google and Bing 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. 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 conduct our topic analysis 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.[5] 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 Enforcement Releases (AAERs) issued over our sample period.[6] This data provides specific details on firms subject to SEC enforcement actions for alleged accounting misstatements.

[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]Consistent with this notion, Dyer, Lang, and Stice-Lawrence [2016] 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.

[6]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.

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).[7] 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. We find that LDA produces a coherent set of economically meaningful topics. These topics capture semantic 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 rela-

[7]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.

tively fluid. The in-sample tests indicate that many of the topic categories are consistently associated with financial misreporting using both the AAER and 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 merger and hedging activities, legal proceedings, and stock option plans. Lastly, our results confirm that the thematic content of annual reports have a significant influence on investor risk perceptions as shown in Bao and Datta [2014].[8]

Our next set of analyses assess the usefulness of *topic* in detecting misstatements compared to a comprehensive set of financial statement and market measures, and textual style characteristics. 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 characteristics include measures of readability, textual complexity, language voice (active and passive), vocabulary variation, emphasis, and disclosure tone.

The out-of-sample tests indicate that *topic* provides significant incremental predictive power over commonly-used financial (*F-score*) and textual style metrics. Specifically, models based on *topic* outperform stand-alone models of *F-score* and textual style features, 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 of *topic* and *F-score* (textual 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. Tests of the predictive 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

[8]Bao and Datta [2014] use LDA to identify the risk types reported in the risk disclosure section (Item 1A) of the annual report. They then correlate these risk topics with stock return volatility over a two month window following the 10-K filing date. We conduct similar tests but instead use all topic categories identified in the 10-K filing.

misstatements. Finally, sensitivity checks indicate that our results are robust to alternative regression specifications, *topic* definitions, corrections for potential model overfitting (Perols et al. [2016]), and restricting our topic analysis to the Management Discussion and Analysis (MD&A) section of firms' 10-K filings(Hoberg and Lewis [2015]).

Our study makes several important contributions to the literature. First, we extend prior research on financial misreporting by documenting that the topics discussed in financial statements 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 portion of corporate disclosures. We exploit a robust textual analysis methodology that directly quantifies *what* is being disclosed in financial statements (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 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, investors, and practitioners, who have begun to implement text-based initiatives aimed at detecting accounting violations. For instance, the SEC is now incorporating basic lists of deceptive words and phrases from corporate filings into its computer-powered detection model, dubbed "Robocop" (Eaglesham [2013]). 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 about what is being disclosed is a rich approach for capturing high-risk accounting activity. 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.[9]

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 and stock market measures.[10] 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

[9]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).
[10]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.

accrual management.

Despite the usefulness of quantitative metrics in detecting accounting manipulations, many argue that the predictive ability of these measures is quite modest, 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 predictive of financial misreporting. Prior studies in this area rely on two general approaches for analyzing text-based disclosures (see Li [2010] and Loughran and McDonald [2016] for 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. 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. Larcker and Zakolyukina [2012] analyze the transcripts of conference calls and find that deceptive language is a better predictor of misreporting compared to abnormal accrual measures.

The second approach employs machine learning algorithms to generate "bags of words" or textual style markers that are predictive of intentional misreporting. Among these studies, Cecchini et al. [2010], 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 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. [2010] uses SVM to generate a dictionary of discriminatory words and phrases that frequently appear in misstated annual reports. They find 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 style

features such as disclosure tone, readability, voice (passive versus active), and lexical variety. This 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.

Our study extends this body of literature by constructing a direct text-based measure aimed at capturing the thematic content of disclosures within firms' financial statements. This approach improves upon prior research by incorporating the deeper semantic meaning of management communications. Further, drawing on theories of deception, we argue that the thematic content, or topics, a manager chooses to discuss is more likely to reflect their intentions, and thus intentional misreporting, than the bag-of-words or textual style approaches used in prior research.

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 reports. The LDA technique is widely-used in the linguistic and information retrieval literature 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] for a brief review), and our measurement approach is consistent with recent studies that use the LDA technique to investigate various 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. [2014] employ LDA topic modeling to compare the thematic content of analyst reports and the text narrative of conference calls. 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 conference calls. Bao and Datta [2014] discover and quantify the various topics discussed in textual risk disclosures from annual 10-K filings (Item 1A). 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.[11]

In a concurrent study, Hoberg and Lewis [2015] use topic modeling and cosine similarity to provide evidence of the content disclosed by firms involved in SEC enforcement actions (AAERs). Focusing on the MD&A section of 10-K filings, Hoberg and Lewis [2015] find that relative to industry peers, AAER firms disclose abnormal verbal content that is common among misreporting firms. Their topic analysis indicates that AAER firms disclose more information about complex business issues such as acquisitions, asset sales, and foreign operations, and are more likely to grandstand good financial performance. AAER firms also provide fewer quantitative details explaining their performance and under-disclose topics related to litigation settlements and liquidity challenges.

Our study differs from Hoberg and Lewis [2015] in several respects. First, Hoberg and Lewis [2015] fit their LDA model using annual reports filed in only the first year of their sample period (1997-2008). This approach does not account for changes in disclosure topics

[11]Of the remaining topics, disclosures of systematic macroeconomic and liquidity risks have an increasing effect on investors' risk perceptions, whereas topics related to diversifiable risks (i.e., human resources, regulatory changes, information security, and operation disruption) lead to a decrease in investors' risk perceptions.

over time and likely induces 'staleness' in the topics used in their empirical analyses. As highlighted in Cecchini et al. [2010] and Li [2010], tracking temporal variations in managerial communication is a much needed extension to current text-based methodologies. We therefore extend Hoberg and Lewis [2015] by accounting for the time-varying nature of management disclosure. Specifically, our methodology simultaneously discovers and quantifies the topics discussed by management in a given reporting year. We also employ a rolling-window estimation procedure that predicts financial misreporting using the topics identified over the five years prior to the manipulation period.

Second, the analysis in Hoberg and Lewis [2015] 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 [2010]), 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 is that companies can strategically shift or (de-)emphasize content across sections.[12] As we will show, topics identified in the MD&A section have lower detection power compared to topics identified from the full annual report. Lastly, and most important, our study goes a step further by demonstrating the incremental predictive power of thematic content beyond commonly-used quantitative and textual style measures. As a result, our study seeks to provide broader insights on the time-varying role of language-based information in detecting financial misreporting, and the potential benefits of statistical topic analysis in assessing the likelihood of financial misreporting.

[12]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. The result potentially reflects managements' intentional obfuscation of negative footnote information with a relatively more positive tone in the MD&A section.

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 and textual style characteristics.

Disclosure topics may provide incremental detection ability since the thematic content of financial statements captures an aspect of managerial deception that is distinct from that of financial metrics. Specifically, regulatory oversight is more difficult for textual narratives, especially at the topic level, thus leaving more room for managers to use disclosure topics as a means of diverting attention away from misstated financials. While prior research has identified a set of "lying words" (see e.g., Newman et al. [2003], Larcker and Zakolyukina [2012]), it is more difficult to naïvely identify a set of "lying topics" as these same topics may be benign or informative about actual performance in other settings or at other points in time. Furthermore, financial metrics in annual reports are primarily backward-looking, whereas textual disclosures contain a significant amount of forward-looking information and cover a wide range of topics (Bozanic, Roulstone, and Van Buskirk [2015]). As prior research suggests, forward-looking information is inherently more uncertain and less verifiable at the time of disclosure, compared to financial statement figures (Bonsall IV, Bozanic, and Merkley [2014]). We therefore explore whether the topics discussed in annual reports provide incremental predictive value beyond financial metrics in detecting financial misreporting. Our first research question is stated as follows:

> **Research Question 1**: *Topic provides predictive information beyond that of financial measures when detecting intentional financial misreporting.*

We also investigate whether disclosure topics provide informational value beyond textual style characteristics in identifying accounting manipulations. This query is warranted as the literature debates whether textual style characteristics reflect managers' intent to deceive or obfuscate information. Textual style characteristics are broad metrics that summarize

various communication patterns found in text documents. These metrics often reflect style features such as language complexity, readability, grammar and word choices, as well as formatting choices such as the use of bullets or the amount of whitespace. Topic modeling, in contrast, captures the underlying context and semantic meaning of discussions within the document.

Prior research on deception suggests that the semantic content of written communication is more intentional relative to style characteristics. Specifically, theories of manipulation and deception suggest that individuals actively monitor the amount, veracity, relevance and clarity of topics communicated (see e.g., McCornack [1992]). Experimental evidence also indicates that individuals adapt deception strategies of giving false answers, withholding relevant information, or giving evasive answers on demand, suggesting that choosing a topic is an intentional process (Buller and Burgoon [1996]). In the context of financial reporting, we expect deceptive managers to consciously select the topics discussed and the attention dedicated to each topic within the financial statements.

While prior studies posit that textual style markers such as length and readability reflect managers' intentional choice to obfuscate or deceive, researchers debate whether this relation simply reflects managers' subconscious slant or the inherent complexity of discussing unusual events or poor performance (see e.g., Li [2008] and the related discussion in Bloomfield [2008]). Further, linguistics research suggests that word choice, in and of itself, is often subconscious (or without intent). For instance, studies find that the use of function words (i.e., pronouns, prepositions, articles, conjunctions, and auxiliary verbs) is often without awareness and is difficult to control (Chung and Pennebaker [2007]). Although studies suggest that individuals consciously choose abstract or concrete words to describe events and behaviors, explicit goals to influence the beliefs of others make it difficult to disentangle whether the communicator is being truthful or not. Thus, even for those words that managers may intentionally choose, it is often difficult to discern intentional deception (see e.g., Douglas and Sutton [2003]). In sum, the above discussion suggests that topic is more likely to reflect

managers' intentional content choices compared to textual style features. This leads to our second research question:

> **Research Question 2**: *Topic provides predictive information beyond that of textual style characteristics when detecting intentional financial misreporting.*

3 Data and Empirical Measures

3.1 Data and Sample Selection

We base out topic analysis on the textual narratives contained in the annual 10-K filings. We focus on annual reports 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.[13] We download the full text of all 10-Ks available through the SEC EDGAR FTP site from January 1, 1994 (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 over the 1994 to 2012 period. We use all the filings to generate the topic measures as this improves the algorithm's convergence. We however exclude all non-U.S. and financial firms (SIC codes 6000-6799) in our prediction analyses.

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.[14] To remove typos and uncommon terminology, we also restrict the words in the files to match those contained in the standard

[13] We acknowledge that 10-Ks 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 not to include quarterly reports to ensure consistency in the disclosure content of the reports across firms' reporting periods.

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

Unix words dictionary.[15] We describe our full parsing methodology in detail in Appendix A.1 of the online appendix. As discussed below, 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.

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 classify firms engaging in material accounting misstatements. We focus on misstatements occurring during the annual reporting period to ensure that the measurement period for our prediction variables is consistent across firms. We create an indicator variable (*misreport*) that equals 1 for each fiscal year identified as misstated by the SEC, and zero otherwise. We then use this indicator variable to classify those 10-K filings containing potential GAAP violations. Our second data source is a customized automated search for occurrences of financial restatements that are seemingly due to intentional misapplications of GAAP (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 financial misreporting.[16]

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

[15]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.

[16]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.

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 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 firm-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 filing does not involve an irregularity.

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).[17] These variables capture accrual quality, firm performance, off-balance-sheet activities, and market-based incentives. We also 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

[17]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. [2010]).

(termed RSST accruals) which captures the change in noncash net operating assets (Richardson et al. [2005]).[18] We also measure the change in receivables and the change in inventory 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 managerial incentives to manipulate their financial statements to mask poor firm performance. 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.[19] Lastly, larger firms are more likely to invest in monitoring mechanisms that

[18]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.

[19]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 leads to qualitatively similar results.

mitigate the occurrence of aggressive accounting activity. We use the log value of total assets to capture potential size effects.

3.2.2 Textual Style Measures

We benchmark our *topic* measure against a comprehensive set of style features used in prior literature, as well as four new measures developed from our analysis. Panel B of Appendix B presents a full list of the style variables and their measurement.[20] Our new measures are the log of the number of bullets, the length of the SEC mandated header, number of excess newlines (vertical whitespace) in the filings, and the character length of HTML tags. The log of the number of bullets captures an aspect of readability, as bulleted information is typically concise. The SEC header contains basic corporate and filing form information such as company name and address, SIC industry, form type, and filing date. Filings with long headers generally identify firms that operated under former corporate names in prior years.[21] We therefore expect the SEC header length to be correlated with complex firm transactions that are correlated with accounting manipulations as noted earlier (e.g., M&As and restructurings).

Excess newlines (vertical whitespace) increase the length of the 10-K filing without adding any substantive content. Deceptive managers could insert additional whitespace to keep the length of the filing consistent with filings by peer firms or the firm's own prior filings while omitting pertinent information. We also include the character length of HTML tags in the unparsed filings to control for structural imprecisions that might be correlated with firm size and time period (Loughran and McDonald [2016]). For instance, SEC filings created using

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

[21]Former company names and the date of each name change are disclosed in a separate block of header fields. Firms can enter up to three former names in the EDGAR system.

less specialized financial software will typically contain numerous HTML tags.

In the vein of Li [2008] and Goel et al. [2010], our next 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 final set of textual 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. Last, we measure disclosure emphasis based on the use of capitalized words, exclamation points, and question marks (Goel and Gangolly [2012]).

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].[22] We choose

[22]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 refrain from using 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.

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.[23] 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. While the accuracy of human classification is greater than that of LDA, the human approach is infeasible when classifying large volumes of textual data. In fact, the LDA tool enables us to categorize the disclosure content of 10-K narratives containing over 3 billion words, allowing for rigorous testing that otherwise would be impossible using human topic classifications.

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, $\beta_K \sim \text{Dirichlet}(\eta)$.[24] The model further assumes that the topic proportions in each document d are drawn from a Dirichlet distribution $\theta_d \sim \text{Dirichlet}(\alpha)$. Given these assumptions, a specific number of topics to identify, and a few learning parameters, the LDA model categorizes the words in a given set of documents into well-defined topics. Because the model uses Bayesian analysis, a word is allowed to be associated with multiple topics. This is a distinguishing feature of LDA, as words can have multiple meanings, especially in different contexts. In 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 we expect disclosure content to change across time due to factors such as macroeconomic condi-

[23]LSA-IDF and LSA-TF are LSA based measures using a term-document matrix that has undergone a transform: inverse document frequency or term frequency, respectively.
[24]A Dirichlet distribution is essentially a multivariate generalization of a beta distribution.

tions, technological changes in business operations, regulatory interventions (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).[25]

Next, we pre-process the parsed 10-K filings by first removing stop words. Stop words

[25] We run the simulation on irregularity restatements given the rarity of SEC AAERs.

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). We compute the topic weights in a given filing 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*.

4 Empirical Results

4.1 Validation of LDA Topic Measure

Before investigating our research questions, we validate our *topic* measure using several methods. Following prior research (e.g., Bao and Datta [2014], Huang et al. [2014]), our first method evaluates the semantic validity of the LDA output by labeling the topics and assessing the extent to which the topics provide meaningful content. 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, 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.[26]

To determine the underlying content of each combined topic, we generate a list of the highest weighted phrases and sentences associated with each topic. 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. 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.[27] The reported bigrams exclude redundant phrases (e.g., "millions in," "company also," "in year") and those with similar inferences (e.g., "compared

[26]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.

[27]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.

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 16 and 54), franchise revenues (topic 26), and general references to quantitative financial statement information (topics 7, 34, 62, and 63). LDA also identifies topics related to complex business transactions and arrangements such as fuel and natural gas purchase commitments (5), derivatives and hedging activities (9 and 41), merger activities (31), R&D partnerships (32), joint ventures (39), strategic alliances (46), and investments in securitized securities (55).

Several topics refer to specific financial statement items and/or their underlying measurement assumptions. These include post-retirement health care cost assumptions (4), account receivables and doubtful accounts (12), long term assets (25), advertising expenses (36), and the measurement of natural gas properties (38). Consistent with Huang et al. [2014], we are able to identify industry-specific topics such as aircraft leasing arrangements in the airline industry, franchise revenue recognition and restaurant growth in the restaurant industry, as well as general discussion 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 (57), country risks (18 and 37), environmental liabilities and risks (6 and 56), patent infringement and rights (48), and legal proceedings (45). In summary, the evidence in Appendix C suggests that the LDA algorithm provides a valid set of semantically meaningful topics.

Our next validation method uses in-sample tests to evaluate whether *topic* performs

reasonably well in detecting misstatements. Figures 1 and 2 depict the distribution of each combined topic over the 1999 to 2012 period (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). 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.[28]

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 (5), and industry-specific topics such as aircraft leasing arrangements (4) and real estate loan operations (10). 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 (39), strategic alliances (46), and partnerships (51) are more prominent in the second half of our prediction period. Likewise, discussions of securitized/guaranteed securities (55) 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 managerial discussion of financial performance is not as clear cut as suggested by prior work linking poor financial performance to accounting manipulations.

[28]Our results are qualitatively similar when we additionally orthogonalize by auditor type (Big N, mid-size, or small).

The results in Figure 1 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 2) 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.

As a final validation of our topic measure, we follow Bao and Datta [2014] and re-examine the relation between disclosure topic and investor risk perceptions as proxied by stock return volatility. We measure return volatility as the standard deviation of daily stock returns over the one-year period following the 10-K filing date. Figure 3 illustrates the association between our combined topics and risk perceptions over time. Similar to Bao and Datta [2014], we find that the thematic content of annual reports is significantly associated with post-disclosure risk perception. However, while Bao and Datta [2014] find insignificant associations for roughly two-thirds of their risk disclosure topics (22 out of 30), our results show that most of the topics generated from the full 10-K filing are significant predictors of future return volatility. Topics that primarily load as positive include fair value/cash flow hedging activities, partnership and joint venture agreements, stock option plans, and securitized/guaranteed securities. Discussions of specific risk factors such as foreign country and currency risks, environmental risks, and fuel costs and commitments are also positively related to future return volatility.

Taken together, our evidence in Figures 1 through 3 suggest that the thematic content of annual reports provides significant informational value for detecting misstatement events as well as predicting investors' risk perceptions. 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.[29] [30] We begin our analyses by estimating 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. For our AAER specification, we lack sufficient missatement events for the years 2011 and 2012, and thus exclude these two years from our out-of-sample 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

[29] 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.

[30] 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.

the financial variables.[31] In both cases, we generate *p_misreport* and run the out-of-sample tests. For RQ2, we introduce four specifications that include style characteristics. The first model includes our textual style metrics with the second including both financial and style characteristics. The third and fourth models are expanded versions of the first two models with *topic* as an additional predictor. Our general regression form for RQ2 is 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 naturally small number of AAERs and irregularity restatements in our test windows. 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.[32] 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

[31] 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.

[32] 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.

across years.

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

Table 1 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 2 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 for AAERs and irregularity restatements, respectively. The statistics indicate that the financial variables (*F-score*) provide a significant amount of information for predicting AAERs ($p < 0.001$). However, they fail to provide significant informational value for predicting irregularity restatements ($p = 0.165$), consistent with the summary statistics reported in Table 1.

Panels B and D present Var-Gamma tests of the predictive ability of *topic* compared to that of *F-score* and a joint specification of both predictors. In the case of AAERs (see Panel B), we find that the stand-alone *topic* vector underperforms the vector of financial metrics ($p < 0.001$). This result is not surprising given that the original Dechow et al. [2011] *F-score* model was build using a sample of AAERs. Nonetheless, the paired vectors of *topic* and *F-score* perform significantly better at predicting AAERs than the stand-alone *F-score* vector ($p < 0.001$). With respect to irregularity restatements, we find that *topic* and the pairing of *topic* with *F-score* both outperform the stand-alone *F-score* specification ($p < 0.001$). We

also find that the predictive ability of *topic* is not significantly different from that of *topic* paired with *F-score* ($p = 0.355$) indicating that financial metrics add little detection power over *topic* for the restatement sample. Overall, these results indicate 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.

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

Table 3 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 (concise) bulleted 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 vectors in the same regression model. Table 4 presents the Fisher and Var-Gamma statistics for out-of-sample tests of the predictive performance of *topic* relative to textual 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 *topic* combined with style is a good predictor of misstatements involving AAERs and irregularity restatements ($p < 0.001$). However, for AAERs, the Var-Gamma results in Panel B show that *topic* by itself is a better predictor of misreporting than either textual style ($p < 0.001$) or *topic* combined with style ($p = 0.038$).

The Var-Gamma tests for irregularity restatements (see Panel D) show that, while style is a better predictor than *topic* ($p = 0.002$), the joint vector of *topic* and style is a better predictor than the stand-alone *topic* and style vectors ($p \leq 0.001$). 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 style characteristics are better able to identify accounting manipulations that do not involve a SEC enforcement action.

4.2.4 Joint Predictive Value of *topic*, Financial, and Textual Style

We conduct extended analyses of the interplay between all three sets of prediction variables: *topic*, financial statement and stock market variables, and textual style characteristics. This comprehensive analysis evaluates whether the predictive ability of *topic* is robust to the inclusion of financial and textual style characteristics. Table 5 presents out-of-sample results from a comprehensive regression of all three vectors of prediction variables. In Panels A and C, we find that the three-vector specification performs reasonably well in detecting accounting misstatements out of sample ($p < 0.001$). 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 ($p < 0.001$). However, the combined vector does not perform any better than the joint vectors of *topic* and financial measures for AAERs ($p = 0.393$ in Panel C) nor the joint vectors of *topic* and textual style measures for irregularity restatements ($p = 0.684$ in Panel D). This evidence corroborates our previous results – *topic* and financial measures are stronger predictors of misstatements involving AAERs, whereas *topic* and textual style provides more robust power for predicting irregularity restatements.

4.3 Economic Significance of *topic* Measure

To gauge the economic significance of our results, we examine the classification accuracy of the out-of-sample regressions 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." Table 6 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 73% of misstatements involving AAERs. The joint model of *topic* paired with *F-Score* performs slightly better flagging 74% of misreported filings, while the three-vector model correctly flags about 75% of misstatements. This evidence corroborates our findings in Table 5. When we focus on high risk prediction scores, we observe that the three-vector model and the joint model of *topic* and *F-score* captures the most misstatements (31.50%) at the 90th percentile. This accuracy rate is 33% higher than the rate achieved using the comparable model of *F-score* paired with textual 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%).

In Panel B, the combined model of *topic* and textual style is the most efficient at capturing irregularity restatements above the 50th percentile. However, this model outperforms the stand-alone style model by only 1 percentage point. For high risk filings, the three-vector model is most accurate, flagging roughly 28% and 17% of misstated filings at the 90th and 95th cut-offs, respectively. When we benchmark the three-vector model to *F-score* paired with textual style, we find that including *topic* as an additional predictor increases detection accuracy by roughly 5-6% at both cut-offs. Taken together, these results suggest that *topic* is an economically significant predictor of financial misreporting and that the incremental value of *topic* is more salient when attempting to identify high risk accounting practices.

To further illustrate the economic significance of 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 in

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.[33]

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.[34] [35]

4.4 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 (errors). Next, we re-estimate *topic* using the MD&A section instead of the full text of the 10-K filings. We also change the regression form to a L1 regularized logit model, to alleviate concerns of potential overfitting. Lastly, we adjust

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

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

[35]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 29), and digital technology and services (combined topic 24).

our samples of misstated filings to exclude repeat GAAP violators as well as 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.

4.4.1 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 all specifications have statistically strong performance for predicting unintentional accounting errors, though *topic* by itself and *topic* paired with F-score perform slightly better than the other models. Given that some errors can escalate to intentional manipulations (perhaps to conceal the error), this evidence suggests that *topic* provides some added value as an early warning sign.

4.4.2 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. [2010], Hoberg and Lewis [2015], 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 Fisher statistics 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 [2010], Loughran and McDonald [2016]).

4.4.3 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. [2016]). 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 tests are similar to our main analyses with one exception – the specification of *topic* paired with financial measures 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.

4.4.4 Removing repeat GAAP Violations

Our main analyses allow a firm to be flagged as a misreporting firm in both the learning window and the following test year. This approach could bias our *topic* measure towards firms that are repeat offenders. To alleviate this concern, we adjust our misstatement samples by removing all misreporting years that are immediately preceded by a misstated firm-year. Thus, our out-of-sample dependent variable only picks up the first year affected by a misstatement. Our prediction results are virtually identical after these deletions.

4.4.5 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.

5 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 prior research on linguistics and deception, we conjecture that this *topic* measure will be 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 traditional financial statement and textual style measures. Specifically, based on out-of-sample tests, we find that models that incorporate *topic* outperform models based on commonly-used financial and style measures. Further, our results reveal that *topic* is 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 narratives contained in 10-K filings are useful in predicting intentional financial misreporting beyond traditionally examined financial measures and 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, practitioners, and investors 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 disclosures, the information they contain, and their effects on markets and the greater economy.

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) Aircraft leasing agreements: commercial aircraft, aircraft engines, boeing aircraft, entered into, aircraft maintenance, operating lease, lease agreement, additional aircraft, aircraft manufacturers, credit corporation
4) 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
5) Fuel costs and commitments: nuclear fuel, fuel related, coal mining, coal reserves, fuel costs, fuel expense, fuel supply, fuel commitments, related expenses, fossil fuel
6) 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
7) Financial statement information: dollars in millions, ended december, income taxes, financial statements, accompanying notes, millions except, year ended, pension benefits, consolidated balance, consolidated financial
8) 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
9) 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
10) Real estate loan operations: real estate, national bank, estate loans, estate investment, loan bank, home loan, federal home, investment trust, trust REIT, mortgage loans
11) Gaming operations: gaming license, gaming operations, mississippi gaming, gaming machines, gaming commission, casino gaming, gaming control, native american, indian gaming, gaming taxes
12) 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
13) 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
14) Mining operations: ounces of gold, gold bank, gold and silver, gold mining, copper gold, mining claims, million ounces, square feet, into gold, gold production
15) Cable/television operations: cable television, television stations, television systems, television industry, cable systems, television programming, cable programming, fiber optic, cable operators, cable act imposed
16) Segment performance: generation segment, discussed below, segment as discussed, because of items, segment because, merchant services, other operations, increased million, items detailed, maintenance expenses
17) Aircraft manufacturing operations: mcdonnell douglas, boeing aircraft, boeing company, commercial aircraft, boeing mcdonell, lockheed martin, agreement between, customers include, savings bank, sales to boeing
18) 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
19) Patient and nursing services: skilled nursing, nursing facilities, assisted living, nursing homes, nursing home, physical occupational, nursing care, living facilities, home health, real estate
20) Marine operations: insurance carriers, marine transportation, marine services, offshore marine, other carriers, licensed insurance, marine containers, maintain insurance, self insure, marine insurance
21) Agricultural operations: crop insurance, crop hail, crop production, agricultural partnerships, crop nutrient, insurance business, crop yields, named peril, crop drying, agricultural market
22) Laser products: excimer laser, laser system, laser vision, laser technology, laser printers, laser based, laser beam, laser products, capital expenditures, discontinued operations

Continued on next page

23) 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
24) 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
25) 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
26) Franchise revenue recognition: franchise fees, franchise agreements, initial franchise, franchise royalties, franchise operations, franchise revenues, franchise rights, development fees, intangible assets, brand name
27) Business structure: holding company, bank holding, loan holding, company under, financial holding, holding corporation, utility holding, holding companies, unrealized holding, holding gains
28) Debt issuance: convertible debenture, common stock, subordinated debenture, agreement with, purchase agreement, principal amount, note or debenture, debenture holders, company issued, debenture offering
29) Media and entertainment: interactive entertainment, entertainment software, entertainment company, entertainment group, agreement dated, entertainment industry, entertainment services, company acquired, lease agreement, digital entertainment
30) 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
31) 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
32) R&D partnerships: pharmaceutical products, collaboration agreement, agreement between, pharmaceutical company, collaboration with, entered into, pharmaceutical services, license agreement, research collaboration, between the company
33) Floral products: crop insurance, crop protection, fresh cut flowers, floral products, specialty retailers, crop nutrient, floral services, named peril, crop year, brooding and weed
34) Consolidated financial information: consolidated statements, cash flows, statements of cash, subsidiaries consolidated, corporation consolidated, consolidated balance, balance sheets, comprehensive income, share amounts, company consolidated
35) 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
36) 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
37) 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
38) 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
39) 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
40) Credit card operations: credit card, debit card, card transactions, card receivables, card services, card processing, gift card, card loans, interchange fees, card revenue,
41) 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
42) Apparel manufacturing: women's apparel, outlet stores, czech republic, apparel group, apparel manufacturers, weekly basis, united states, business segments, dominican republic, mens apparel
43) Purchase agreements: common units, each purchaser, authorized purchaser, units purchased, agreement with, with the purchaser, entered into, third party, affiliated purchaser, place an order
44) 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
45) 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
46) Strategic alliances: strategic alliance, alliance with, entered into, agreement with, alliance agreement, alliance partners, into a strategic, company entered, license agreement, ventures sold

Continued on next page

47) Credit agreements: jpmorgan chase, kinder morgan, vice president, agreement with, entered into, credit facility, facility with, chase manhattan, credit agreement, company entered
48) 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
49) 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
50) Share capital: preferred stock, redeemable preferred, mandatorily redeemable, cumulative redeemable, investing activities, convertible preferred, series A cumulative, common stock, series A preferred, capital securities
51) Partnership arrangements: general partner, limited partner, sole general, managing general, operating partnership, managing partner, limited partnerships, partner interest, executive officers, responsible for managing
52) Equity ownership and control: life insurance, common stock, consolidated statements, property trust, insurance company, agreement with, limited partnership, property limited, year ended, december ended
53) 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
54) Segment performance: segment information, operating segment, business segment, segment consists, operating income, segment performance, performance based, reportable segment, segment reporting, segment results
55) 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
56) Environmental risks: mining operations, environmental liabilities, environmental remediation, environmental regulation, environmental risks, environmental laws, environmental compliance, environmental protection, environmental costs, mining claims
57) 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
58) 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
59) Short-term credit facilities: credit facility, revolving credit, senior secured, entered into, secured credit, senior credit, term loan, credit agreement, secured revolving, term debt
60) 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
61) 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
62) 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
63) 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
64) 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

This table presents summary information for each of the combined topics. The topic number and label are presented in bold. The right side is based on a set of "representative sentences." These sentences were constructed by finding the top 1000 sentences in terms of topic weighting per length (excluding stop words). Then, the middle 334 in terms of length were selected. The right side shows the 10 most common n-grams, snippets of text between words of at least four characters.

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.

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.

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.

BOZANIC, Z., D. T. ROULSTONE, and A. VAN BUSKIRK. 'Attributes of Informative Disclosures.' Working paper, Ohio State University. 2015.

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.

BULLER, D. B. and J. K. BURGOON. 'Interpersonal deception theory.' Communication Theory 6 (1996): 203–242.

BUSHEE, B. J., I. D. GOW, and D. J. TAYLOR. 'Linguistic Complexity in Firm Disclosures: Obfuscation or Information?' Working paper, University of Pennsylvania. 2015.

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 (2010): 164–175.

CHUNG, C. and J. W. PENNEBAKER. 'The psychological functions of function words.' Social Communication (2007): 343–359.

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 ever-expanding 10-K: Why are 10-Ks getting so much longer (and does it matter)?' Working paper, University of North Carolina – Chapel Hill. 2016.

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

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

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.

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 Engage in Disclosure Herding?' Working paper, University of Southern California. 2015.

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.' Working paper, Hong Kong University of Science and Technology. 2014.

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.

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

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

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.

MCCORNACK, S. A. 'Information manipulation theory.' Communications Monographs 59 (1992): 1–16.

MCLEAN, B. 'Is Enron Overpriced? It's in a bunch of complex businesses. Its financial statements are nearly impenetrable. So why is Enron trading at such a huge multiple?' Fortune Magazine (Mar. 2001).

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.

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, In-Press. 2016.

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.

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.

Figure 1: Combined Topic Distribution and Irregularity Restatement Prediction

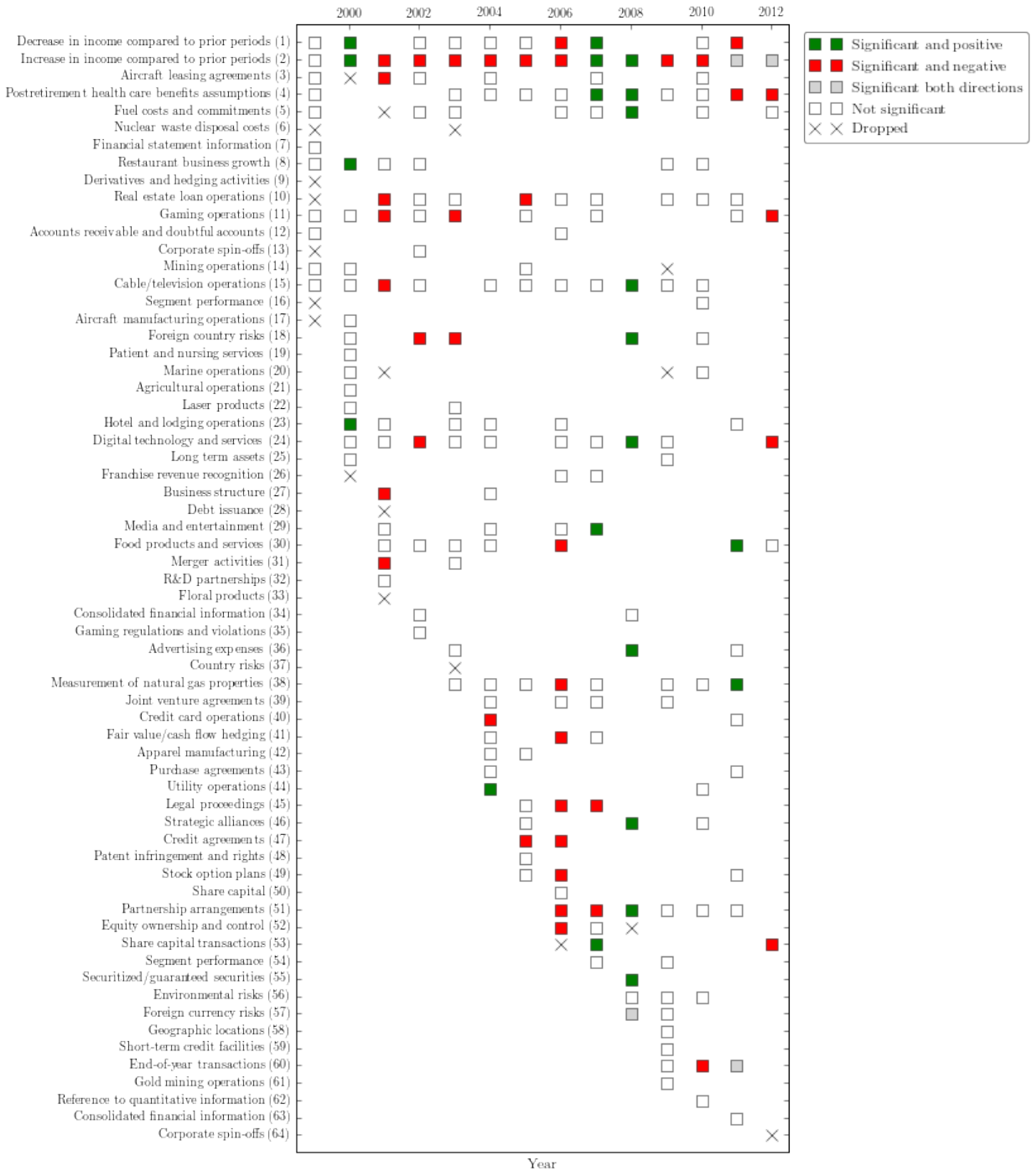

Figure 1: Combined Topic Distribution and Irregularity Restatement Prediction (Continued)

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 (Continued)s). White square boxes indicate that the disaggregated subtopics are insignificant in that particular year.

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.

Figure 3: Combined Topic Distribution and Investor Risk Perceptions

Figure 3: Combined Topic Distribution and Investor Risk Perceptions (Continued)

This figure illustrates the association between each combined topic and investor risk perceptions as captured by stock return volatility over the one-year period following the 10-K filing date. We estimate yearly OLS regressions of stock return volatility on the disaggregated subtopics associated with each combined topic. 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 dropped from our regressions due to collinearity are denoted as an X. The square boxes are color coded based on the direction and statistical significance of the disaggregated subtopics in a given year. The square boxes are color coded green (red) if the subtopics are positively (negatively) related to investor risk perceptions. The box is color coded grey if the subtopics are statistically significant but in opposing directions. White square boxes indicate that the disaggregated subtopics are insignificant in the respective year.

Table 1: 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 2: Out-of-Sample Prediction Analysis of *topic* and *F-score*

Panel A: Fisher Tests (AAER Sample)

Specification	Fisher Statistic	p-value
F-score	151.94^{***}	< 0.001
topic	113.90^{***}	< 0.001
topic and *F-score*	194.35^{***}	< 0.001

Panel B: Var-Gamma Tests (AAER Sample)

	F-score	*topic* and *F-score*
topic	-38.04^{***}	-80.45^{***}
	(< 0.001)	(< 0.001)
topic and *F-score*	42.41^{***}	
	(< 0.001)	

Panel C: Fisher Tests (Irregularity Restatement Sample)

Specification	Fisher Statistic	p-value
F-score	35.19	0.165
topic	95.53^{***}	< 0.001
topic and *F-score*	85.84^{***}	< 0.001

Panel D: Var-Gamma Tests (Irregularity Restatement Sample)

	F-score	*topic* and *F-score*
topic	60.35^{***}	9.69
	(< 0.001)	(0.355)
topic and *F-score*	50.66^{***}	
	(< 0.001)	

This table provides comparative out-of-sample tests of our prediction models. Section 3.2.3 provides a detailed discussion of the construction of the *topic* measure. Appendix B defines the financial variables (denoted as *F-Score*) used in our prediction models. Panels A and C present test statistics using using Fisher's [1932] method 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 from the out-of-sample predictions generated by regressions of *misreport* on $p_{misreport}$ generated by the rolling five-year windows. Degrees of freedom for the tests in panels A and C are 24 and 28, respectively. Panels B and D present the Var-Gamma statistical tests (see Appendix D) for the AAER and irregularity restatement samples using λ equal to 12 and 14, respectively. 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 3: 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
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^{***}
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										
$Coleman\text{-}LiauIndex$	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

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 4: Out-of-Sample Prediction Analysis for *topic* and *Style*

Panel A: Fisher Tests (AAER Sample)

Specification	Fisher Statistic	p-value
Style	34.44^{*}	0.077
topic	113.90^{***}	< 0.001
topic and *Style*	93.44^{***}	< 0.001

Panel B: Var-Gamma Tests (AAER Sample)

	Style	*topic* and *Style*
topic	79.47^{***} (< 0.001)	20.47^{**} (0.038)
topic and *Style*	59.00^{***} (< 0.001)	

Panel C: Fisher Tests (Irregularity Restatement Sample)

Specification	Fisher Statistic	p-value
Style	128.42^{***}	< 0.001
topic	95.53^{***}	< 0.001
topic and *Style*	163.37^{***}	< 0.001

Panel D: Var-Gamma Tests (Irregularity Restatement Sample)

	Style	*topic* and *Style*
topic	-32.89^{***} (0.002)	-67.84^{***} (< 0.001)
topic and *Style*	34.94^{***} (0.001)	

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 for the AAER and irregularity restatement samples, respectively (see Section 3.1.1 for a description our data sources). The test statistics are based on an aggregation of the p-values from the out-of-sample predictions generated by regressions of *misreport* on $p_{misreport}$ generated by the rolling five-year windows. Degrees of freedom for the tests in panels A and C are 24 and 28, respectively. Panels B and D present the Var-Gamma statistical tests (see Appendix D) for the AAER and irregularity restatement samples using λ of 12 and 14, respectively. 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 5: Out-of-Sample Prediction Analysis for *topic*, *F-score*, and *Style*

Panel A: Fisher Tests (AAER Sample)

Specification	Fisher Statistic	p-value
F-score and *Style*	164.44^{***}	< 0.001
topic, *F-score*, and *Style*	202.62^{***}	< 0.001
topic	113.90^{***}	< 0.001
topic and *F-Score*	194.35^{***}	< 0.001

Panel B: Var-Gamma Tests (AAER Sample)

	F-score and *Style*	*topic*, *F-score*, and *Style*
topic	-50.54^{***}	-88.71^{***}
	(< 0.001)	(< 0.001)
topic and *F-Score*	29.91^{***}	-8.27
	(0.003)	(< 0.393)
topic, *F-score*, and *Style*	38.17^{***}	
	(< 0.001)	

Panel C: Fisher Tests (Irregularity Restatement Sample)

Specification	Fisher Statistic	p-value
F-score and *Style*	141.46^{***}	< 0.001
topic, *F-score*, and *style*	167.62^{***}	< 0.001
topic	95.53^{***}	< 0.001
topic and *Style*	163.37^{***}	< 0.001

Panel D: Var-Gamma Tests (Irregularity Restatement Sample)

	F-score and *Style*	*topic*, *F-score*, and *Style*
topic	-45.92^{***}	-72.08^{***}
	(< 0.001)	(< 0.001)
topic and *Style*	21.91^{**}	-4.25
	(0.040)	(0.684)
topic, *F-score*, and *Style*	26.16^{**}	
	(0.015)	

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 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 from the out-of-sample predictions generated by regressions of *misreport* on $p_{misreport}$ generated by the rolling five-year windows. Degrees of freedom for the tests in panels A and C are 24 and 28, respectively. Panels B and D present the Var-Gamma statistical tests (see Appendix D) for the AAER and irregularity restatement samples using λ of 12 and 14, respectively. 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: Detection Performance of *topic* for AAERs and Irregularity Restatements

Panel A: AAER Detection

	50th	90th	95th
topic	72.54	18.60	11.25
F-score	71.16	23.86	14.04
Style	60.21	11.95	6.50
topic and *F-score*	74.07	32.07	17.24
topic and *Style*	74.47	19.40	11.27
F-score and *Style*	73.98	23.73	14.66
topic, *F-score*, and *Style*	75.09	31.50	21.44

Panel B: Irregularity Restatement Detection

	50th	90th	95th
topic	65.19	19.74	10.49
F-score	58.33	15.96	9.95
Style	69.68	25.54	14.06
topic and *F-score*	63.31	21.40	11.30
topic and *Style*	70.61	24.55	15.56
F-score and *Style*	69.60	26.72	15.94
topic, *F-score*, and *Style*	70.18	28.05	16.83

This table reports the classification accuracy of our detection models using out-of-sample prediction scores above the 50th, 90th, and 95th percentile of the sample distribution in each year. Panel A (Panel B) reports the results for the AAER (irregularity restatement) sample. For each prediction model, we report the average annual percentage of misstated 10-K filings that are accurately classified as misstated at the respective percentile cut-off.