Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection*

LYNNETTE PURDA, *Queen's University*

DAVID SKILLICORN, *Queen's University*

1. Introduction

A recent survey by Ernst & Young (2010) suggests that fraudulent activity has increased in the postfinancial crisis years while at the same time resources allocated to fight fraud have been cut. As a result, those responsible for detecting fraud, including auditors, regulators and investors, must be judicious in choosing cost-effective tools to help them with this task. In this paper we add to the set of fraud-detection tools by developing a statistical method for analyzing the language used in the management discussion and analysis (MD&A) section of a firm's annual and quarterly reports. While we show this method to be a highly effective predictor of fraud, the contributions of this paper go further than the development of a tool alone. A key objective of our work is to inform and assist those responsible for choosing between alternative fraud-detection techniques by providing a thorough assessment of financial, language-based, and nonfinancial fraud-detection tools. Our goal is not only to establish which methods are effective predictors of fraud but also to provide the correlation between the various approaches, so that a decision-maker has some basis for choosing amongst the various tools and can eliminate those that unnecessarily increase cost or duplicate efforts.

The language-based tool that we develop relies on the data to identify what is and is not an important indicator of fraud. Similar to the statistical approach used by Li (2010a) to classify the tone of a report, this approach requires us to "train" the tool to identify markers of fraud and truth. To do so, we construct a sample of quarterly and annual reports issued by firms with at least one Accounting and Auditing Enforcement Release (AAER) asserting fraud. Since we make no previous assumptions about what is or is not indicative of fraud, we use these releases to define fraudulent versus truthful reports and rely on the underlying reports to generate fraudulent markers. We do so by using a decision-tree approach to establish a rank-ordered list of words from the MD&A sections that are best able to distinguish between fraudulent and truthful reports. Based on the top 200 words from this list, we use support vector machines (SVMs) as in Antweiler and Frank (2004), Cecchini, Aytug, Koehler, and Pathak (2010), and Humphreys, Moffit, Burns, Burgoon, and Felix (2011) to predict the status of each report and assign it a probability of truth. An advantage of this approach is that we do not require any previous knowledge of what may constitute a suspicious word and it can be easily updated on the basis of new reports, thereby answering Zhou and Kapoor's (2011) call for adaptive and evolutionary fraud-detection techniques. Across all reports from this sample, we find the rate of correct classification to be approximately 82 percent.

* Accepted by Theresa Libby. The authors would like to thank the Social Sciences and Humanities Research Council (SSHRC) and the CPA-Queen's School of Business Centre for Governance for helpful funding for this project.

doi:10.1111/1911-3846.12089

Having established that our data-generated approach can produce strong classification results in isolation, we proceed to compare it to other fraud-detection tools across a variety of samples. The specific fraud-detection methods we examine are: Dechow, Ge, Larson, and Sloan's (2011) *F*-score; capacity differentials indicating a mismatch between revenue growth and the use of firm resources (Brazel, Jones, and Zimbelman 2009); unexplained audit fees (Hribar, Kravet, and Wilson 2010); involvement in merger and acquisition (M&A) activity; and four predefined language dictionaries related to deception, negativity, uncertainty, and litigation. We choose these approaches because they represent a variety of financial and nonfinancial detection methods and have been shown in previous work to be both effective and at most moderately correlated with one another (e.g., Brazel et al. 2009; Price, Sharp, and Wood 2011).

We follow previous studies by Cecchini et al. (2010), Goel, Gangolly, Faerman, and Uzuner (2010), Humphreys et al. (2011), and others by conducting the comparison of fraud detection methods across a cross-sectional sample of fraudulent and truthful *annual* reports. In justifying a focus on annual reports alone, Brazel et al. (2009) emphasize the need to be consistent with other studies, the unaudited nature of interim reports leading to differences in disclosure, and the reduced likelihood that discrepancies in financial reporting will appear in a shorter, quarterly time horizon. We note, however, that excluding fraudulent quarterly reports reduces the observed number of frauds by approximately 24 percent for Brazel et al. (2009) and 17 percent for Dechow et al. (2011). Clearly quarterly misstatements represent a significant number of frauds that are important for both auditors and investors to detect.

We argue that linguistic analysis may be particularly well-suited to identifying unusual discrepancies in financial reporting over shorter quarterly time horizons, consistent with Goel and Gangolly's (2012, 76) suggestion that "the quantitative financial numbers contain redundant information that does not change when a company is committing fraud but the writing style and the presentation style employed to communicate financial information changes." As a result, we examine the effectiveness of the five language-based approaches we study in identifying truth from fraud in a time-series sample of quarterly and annual reports. This is a challenging context for any detection tool given the similarities in a firm's reporting style from period to period and the potential use of standardized or boilerplate language (Brown and Tucker 2011). However this is precisely the environment within which accounting professionals practice as they seek to determine when a firm moves from truthful reporting to misrepresentation and fraud.

In comparing the performance of our data-generated word list to other predefined word lists on a time-series sample of 10-Q and 10-K reports we provide two additional contributions. First we show that language can be used effectively to identify fraud in a time-series setting containing interim reports. It is particulary important to know that word choice may raise red flags given Kaminski, Wetzel, and Guan's (2004) and Dechow et al.'s (2011) findings that a firm's financial numbers show few significant differences during fraudulent periods compared to surrounding truthful reports. Despite the lack of difference, Abbassi, Albrecht, Vance, and Hansen (2012) find that the inclusion of quarterly financial ratios improves fraud detection rates beyond what can be achieved by examining annual figures alone and our results confirm the usefulness of interim reports for a language-based approach as well.

Our second significant contribution related to the inclusion of quarterly reports is to show that the language used in a firm's own previous reports serves an important benchmarking function against which to identify fraud for that particular firm. We find that from the quarter directly preceding the fraud to the first fraudulent report there is a dramatic and significant drop in the probability of truth assigned to a firm's report. Including a measure of the *change* in the probabilty-of-truth variable from one quarter

to the next provides incremental predictive power in identifying fraud for the firm. In other words, keeping the probability of truth fixed at a given level, knowing the degree and direction of change from the previous report provides additional predictive power. These results answer previous calls to extend the literature by including a temporal component tracking language use (e.g., Cecchini et al. 2010; Li 2010b) and suggest that, much as auditors rely on a firm's prior trends in financial performance to identify deviations from the norm, language-based tools will also benefit from the use of firm-specific benchmarks.

In the next section of the paper, we provide an overview of the existing literature on textual analysis, highlighting previous work on fraud detection. We emphasize where previous work has been both similar to and different from our own. Section 3 of the paper describes our data-generated approach to fraud detection and provides an overview of its results. Section 4 introduces the alternative fraud detection methods and conducts a comparison of their effectiveness and correlation across various samples of 10-K reports. Section 5 examines the use of language-based tools for detecting fraud in both annual and interim reports. Section 6 concludes by summarizing the major contributions of our work and its implications for the accounting profession.

2. Textual analysis of corporate disclosures and fraud detection

Textual analysis has become an increasingly popular tool for examining the information content of corporate disclosures. Language-based tools have been used to examine executive conference calls (Larcker and Zakolyukina 2012), earnings press releases (Davis, Piger, and Sedor 2012) and, of course, annual reports (Loughran and McDonald 2011; Li 2008, 2010a; Brown and Tucker 2011). As the academic literature applying textual analysis to accounting and finance contexts has grown, so too have signs of its relevance to practitioners. On the finance side, there have been the launch of several companies analyzing the language used in corporate disclosures, the press, and social media to track market sentiment.[1] On the accounting side, the Big 4 audit firm Ernst & Young has recently developed fraud-monitoring software capable of searching the language used in employee emails for signs of corporate malfeasance (Donahue 2013) while the SEC is developing software to examine language use in financial reports for signs of fraud (Eaglesham 2013).

Within the analysis of language used in annual reports lies a large number of studies focusing on the MD&A section specifically (Feldman, Govindaraj, Livnat, and Segal 2010; Cecchini et al. 2010; Humphreys et al. 2011). The MD&A has been earmarked as being especially relevant since it provides investors with the opportunity to see the company's performance and prospects through management's eyes. Humphreys et al. (2011, 587) suggest that the MD&A is the most read section of annual reports and is particularly useful for textual analysis in that it "does allow for more diverse and less prescribed language than other parts of the 10-K."

From a practical perspective, the relevance of the MD&A can be seen from changes brought about by the Sarbanes-Oxley Act demanding enhanced managerial discussion in this portion of financial reports (Li 2010b). Subsequent to these changes, the SEC issued a 2003 release providing detailed guidance on drafting the MD&A section and launched a 2005 civil lawsuit against the executives of Kmart claiming inadequacies in their MD&A.[2] The language analysis tool currently under development by the SEC is targeted at the MD&A section specifically. External auditors are also paying more attention to this

1. Examples of firms tracking market sentiment or corporate disclosures through textual analysis include FINIF Financial Informatics, SNTMNT, and Lexalytics.
2. A discussion of the SEC's guidelines, its actions against Kmart Corp and advice for CPAs involved in drafting the MD&A can be found in Meiers (2006).

section as shown by a 2012 report from the Canadian Public Accountability Board suggesting enhanced auditor assurances related to parts of the MD&A.[3]

As described in Li's (2010b) survey of the literature, studies analyzing the language used in corporate reports generally take one of two approaches. The first approach draws heavily on previous work in linguistics and psychology and relies on predefined lists of words ("bags of words") thought to be associated with a particular sentiment such as negativity, optimisim, deceptiveness, or uncertainty. As Loughran and McDonald (2011) emphasize, word lists designed to capture these sentiments in ordinary speech may not apply perfectly to business documents. They show that by adapting lists to the financial context, associations between the appearance of these words and stock price movements can be found. In the context of financial statement misrepresentation specifically, they find an association between their Fin-Negative, Fin-Uncertain and Fin-Litigious word lists and instances of fraud.

Other studies relying on "bag of word" methods to detect fraud include Larcker and Zakolyukina's (2012) examination of word categories indicating anger, anxiety, and negation defined by the Linguistic Inquiry and Word Count dictionary (Pennebaker, Chung, Ireland, Gonzales, and Booth 2007). As part of their analysis, Goel et al. (2010) look to distinguish truthful from fraudulent reports on the basis of the number words corresponding to uncertain, positive, and negative word lists. Humphreys et al. (2011) also search for uncertainty by examining the proportion of modal verbs used in reports, hypothesizing that words indicative of uncertainty are significantly higher in deceptive versus truthful statements.

The second approach used to analyze language relies on statistical methods to allow the data to identify what is or is not an interesting word. While this approach has its origin in computer science, it has crossed into business research with the use of these methods to classify the tone of Internet postings (Antweiler and Frank 2004) and forward-looking statements in 10-K and 10-Q reports (Li 2010a). Li (2010b) argues that this approach has several advantages over bag of word methods including the fact that no adaptation to the business context is required.

Given its advantages, statistical methods have been used to detect fraudulent financial statements by several researchers including Goel et al. (2010), Cecchini et al. (2010), Humphreys et al. (2011), Glancy and Yadav (2011), and Goel and Gangolly (2012). In all of these studies, a matched sample of fraudulent and non-fraudulent annual reports is created and, with the exception of Goel et al. (2010); the MD&A section extracted.[4] Among these papers, Cecchini et al. (2010) and Goel et al. (2010) are closest to our work; however, they are similar to us in very different ways. Goel et al. (2010) examine a host of different linguistic features as predictors of fraud but are similar to us in their choice to include some annual reports from fraudulent firms both before and after the period involving fraud rather than the fraud period alone. Our paper extends this analysis by including both quarterly and annual reports from fraud firms in time series with the goal of identifying deviations in a firm's benchmark style as a possible indicator of fraud. Goel et al. (2010) acknowledge the absence of quarterly reports as a limitation of their work while Cecchini et al. (2010, 174) speak more broadly about the limitations associated with purely cross-sectional analysis of language. They state quite clearly that "A possible addition to the methodology would be the inclusion of a temporal component to track

3. This report is available at http://www.cpab-ccrc.ca/EN/content/CPAB_Public_Report_2012_Eng.pdf accessed May 30, 2013.

4. Goel and Gangolly (2012) conduct their analysis using both the entire text of the 10-K report and the MD&A section in isolation.

changes to the structure of text over time for individual firms. Such an extension would allow researchers to better understand how firm changes are manifest in financial texts."

While Cecchini et al. (2010) do not examine a time series of reports, their work is similar to ours in that they compare a language-based method of fraud detection to financial analysis tools more commonly used by auditors and other practitioners. Their comparison is confined to the ratios identified by Beneish (1999) as indicators of fraud whereas we compare the performance of Dechow et al.'s (2011) *F*-score, capacity differentials (Brazel et al. 2009), unexplained audit fees (Hribar et al. 2010), and involvement in M&A activity to five language-based indicators. As a result, our work seeks to both broaden comparisons of qualitative and quantitative fraud-detection tools and examine the possible benefits of highlighting the change in a firm's language use from one reporting period to the next in identifying financial misstatements.[5]

3. Data-generated language tool

Sample construction and methodology

To create a data-generated word list capable of distinguishing fraud from truth, we refer to the SEC's AAER bulletins to identify firms with financial misrepresentations. We search each bulletin issued between October 1999 and March 2009 (bulletin numbers 1190 to 2762) for the words "fraud," "fraudulent," "anti-fraud," and "defraud." In a manual review of a sample of the resulting bulletins, we find that the SEC is very direct in its use of these words and our method provides an effective screen. Examples of phrases found in the bulletins include "fraudulently decided to eliminate" (Release 2108) and "Every one of the financial statements set forth above was fraudulent" (Release 1422). In both cases, the SEC continues to describe actions as fraudulent despite receiving and accepting an Offer of Settlement from the respondents that neither admitted nor denied the findings of the bulletin.

After removing firms from the finance industry[6] and those without coverage in the COMPUSTAT database we are left with 240 firms with at least one AAER bulletin asserting fraud.[7] For these firms, we record the precise quarters and financial reports identified by the SEC as being fraudulent and download all available 10-Q and 10-K reports on the EDGAR database between 1994 and 2006.[8] We begin in 1994 since this roughly coincides with the launch of EDGAR and proves to be the first year with a reliable number of electronically available reports. We end with the year 2006 since the average number of quarters in our sample between a fraud occurring and the release of a relevant AAER bulletin is 13.4 or over three years. As a result, the AAER bulletins we collect in 2009, the end of

5. While Brown and Tucker (2011) find that changing language indicators in the MD&A can provide valuable information to stock market participants, and the degree of similarity between subsequent firm disclosures has been examined in the context of IPO prospectuses (Hanley and Hoberg 2012), we are not aware of any previous papers that have acknowledged the possibility of changing characteristics of financial reporting being indicative of fraud.

6. There is no consistent approach to the exclusion of financial firms. For instance, Dechow et al. (2011) opt to include firms in this industry when examining fraud while Beneish (1997) excludes them. Our decision for exclusion is based on the case made by Cole and Jones (2005) that the SEC's industry guides require particular disclosures for real estate partnerships, oil and gas, mining, property and casualty insurance, and bank holding companies. By excluding firms with two-digit SIC codes between 60 and 67 we eliminate firms in the real estate, banking, and insurance industries. The remaining industries identified to have unique disclosures make up a trivial portion of our sample at less than 1.6 percent of observations.

7. The most comprehensive study using AAER bulletins to compile a list of fraudulent firms is Dechow et al. (2011) who identify a sample of 350 firms beginning in 1982. Since we are restricted to the electronic availability of financial reports, our sample period begins much later.

8. We ignore 10-KSB and 10-QSB reports that smaller firms may choose to use for their reporting purposes and exclude reports with MD&A sections that are atypically small (5KB or less) which generally incorporate the MD&A section by reference to another filing by the firm.

our sample period, typically refer to misrepresentations occurring in 2006 or earlier fiscal years. The resulting data set consists of 4,895 10-K and 10-Q reports, of which approximately 23 percent are fraudulent.[9]

For each of the sample reports, we extract the MD&A section and keep all text associated with it, including table captions. The only text we choose to eliminate is statements whose primary function is legal rather than informative. Specifically, we exclude legal disclaimers appearing in the Forward Looking Statements section of the MD&A since their style of speech is very different from the rest of the report.[10] Given that these legal disclosures generally consist of a single paragraph that is similar across all firms, they are unlikely to contribute significantly to distinguishing between truth and fraud.

Our fraud-detection technique involves three steps that are described in brief here. Extensive details on the precise tools, software and parameters used can be found in Appendix 1. The first step is common to bag of word methodologies and involves creating a table of word frequencies to identify all words appearing in the sample MD&A sections. We allow for different uses of a word to be counted separately. For instance the word "market" can be used either as a noun to indicate a place to sell goods or as a verb indicating promotion and each usage is counted separately.[11] The frequency of appearance of the various words is standardized for the length of the MD&A section in which it is found so that higher word counts do not simply correspond to longer reports.

The second step in our methodology is to use a decision tree-based approach called Random Forests (Breiman 2001) to sort the words in rank order from most to least predictive. At this stage, 25 percent of sample reports are set aside and reserved for out-of-sample tests and 3,000 individual decision trees are grown in the following way. A bootstrapped sample of MD&A sections is selected at random with replacement for each tree. At each internal node of the tree, the goal is to create an inequality based on one of the words (for example, frequency of word i is < 5) which involves selecting both a particular word and a value to which to compare its frequency. In contrast to word frequency scoring methods, where relationships between the appearance of words and a particular sentiment or report status are generally monotonic such that greater or lower proportionate frequencies of a word are associated with characteristics such as optimism or deception, the relationships chosen here need not be. Relationships may be more subtle so that high frequency word usage may not be the characteristic that identifies the report to be fraudulent. This makes it difficult for potential fraudsters to avoid detection by simply knowing which words are most predictive.

9. By construction, the concentration of fraudulent observations is high; however, it is not much higher than the proportion of firms with nontrivial accounting restatements, which Larcker and Zakolyukina (2012) find exceeds 20 percent in some years of their sample nor is it as high as that occurring in papers employing a one-to-one matching methodology between fraudulent and truthful reports (Cecchini et al. 2010; Goel et al. 2010; Humphreys et al. 2011).

10. An example of an excluded legal statement is the following from Aura System Inc.'s 2005 Q1 report: "This Form 10-Q report may contain forward-looking statements which involve risks and uncertainties. Such forward-looking statements include, but are not limited to, statements regarding future events and the Company's plans and expectations. The Company's actual results may differ significantly from the results discussed in forward-looking statements as a result of certain factors, including those discussed in the Company's Form 10-K for the period ended February 29, 2004, as amended, and this report. The Company expressly disclaims any obligations or undertaking to release publicly any updates or revisions to any forward-looking statements contained herein to reflect any change in the Company's expectations or any events, conditions or circumstances on which any such statement is based. This report includes product names, trade names and marks of companies other than the Company. All such company or product names are trademarks, registered trademarks, trade names or marks of their respective owners and are not the property of the Company."

11. We employ Creasor and Skillicorn's (2012) Q-Tagger tool to identify separate instances of the same word being used differently.

The word frequency chosen at each tree node should be the one that best separates truthful and fraudulent reports. Whenever a word is chosen over its peers because its predictive or discriminative power at the current node is strongest, this is recorded. When all of the trees have been constructed, the words can be ranked according to how often they were used during construction. This ranking is a robust estimate of their global predictive power.

Appendix 2 provides the rank-ordered list of most predictive words. In some cases it is clear that the words correspond to particular events in a firm's life that have previously been shown to be associated with fraudulent activity. For instance, words related to merger activity (e.g., acquisition, acquired), potential legal problems (e.g., settlement, legal, judgments), and financing activities (e.g., debt, lease) appear on the list. However other words present such as "customers," "fiscal," or "gross" seem completely innocuous—it is variations in the frequencies of these words that are predictive—and a company would be hard-pressed to write an MD&A section without referring to these terms.

On the basis of the top 200 predictive words, the third step in our procedure uses SVMs to predict the probability of each report being truthful. We train multiple SVMs using various portions of the data so that more than one prediction of truthfulness can be generated for each report. Ultimately, we average these predictions to generate a single probability-of-truth measure for each MD&A section.

Classification rates

Panel A of Table 1 provides summary statistics on the resulting probabilities of truth that emerge from this method for our sample of 4,895 interim and annual reports. Our first observation from this summary is that, overall, financial reports have a relatively high probability of truth, with a mean value of 80 percent and an even higher median value of 88 percent. Looking at these same values separately for truthful versus fraudulent reports demonstrates the potential usefulness of this method as a classifier. Truthful reports receive an average probability of truth of 87 percent in contrast to only 56 percent for fraudulent reports. Not only do the fraudulent reports receive significantly lower probabilities of truth; the standard deviation of these probabilities is far greater at 0.23 in comparison to 0.11 for truthful reports.

To use this approach to classify reports into truthful versus fraudulent, we need to choose a boundary value for the probability of truth below which firms are predicted to be fraudulent. To some extent, the choice of this boundary is arbitrary. Much like the *F*-score, for which Dechow et al. (2011) decide that a value exceeding 1 predicts fraud, setting this boundary requires a trade-off between correctly identifying a high proportion of frauds versus generating large numbers of false positives, in which truthful reports are mistakenly classified as fraudulent. In panel B of Table 1, we present classification rates based on a boundary set at the mean level of probability of truth across the entire sample (80 percent). However, we also present the area under the Receiver Operating Characteristic (ROC) curve as a more complete measure of classification power, as in Larcker and Zakolyukina (2012) and Price et al. (2011).[12] A ROC area of 0.50 implies that a characteristic is no better than a random classifier while measures above this value indicate predictive power.

Panel B provides classification rates for both the full sample and the 25 percent of reports that were not used to train the data to identify markers of fraud versus truth. We see that, for both the full and withheld sample, overall correct classification rates, that is

12. Since estimation of the ROC curve requires that higher values of a measure be associated with greater predictive ability, we alter our probability-of-truth measure to be the probability of fraud or simply 1-probability of truth.

TABLE 1
Summary statistics and classification results

Panel A: Summary statistics on probability of truthful reporting

	Mean	Median	SD
Full sample	0.80	0.88	0.20
Truthful reports	0.87	0.89	0.11
Fraudulent reports	0.56	0.57	0.23
T and Chi-Square stat for test of difference	61.99	1,000	
(*p*-value)	(0.00)	(0.00)	

Panel B: Predicted versus actual status for the full sample and withheld reports with 80 percent as the fraud threshold

All reports				Withheld reports Not used to discover predictive words			
	Predicted				Predicted		
Actual	Non-fraud	Fraud	Total	Actual	Non-fraud	Fraud	Total
Non-fraud	3,119	649	3,768	Non-fraud	786	154	940
Fraud	223	904	1,127	Fraud	54	226	280
Non-fraud	82.78%	17.22%	100%	Non-fraud	83.62%	16.38%	100%
Fraud	19.79%	80.21%	100%	Fraud	19.29%	80.71%	100%
Correct classification	82.19%			Correct classification	82.95%		
ROC area	0.89			ROC area	0.89		

Notes:

Panel A of the table provides summary statistics on the probability-of-truth measure for reports based on our textual analysis of the MD&A section of annual and interim reports. We show the mean and median probability measures for the entire sample of 4,895 reports and separately for those that are truthful versus those that were subsequently identified to be fraudulent. We see that fraudulent reports are assigned a significantly lower probability of truth. Panel B of the table provides classification matrices for how well these probabilities can correctly identify truthful and fraudulent reports. Results are provided both for the entire sample of reports, and those withheld from the sample during the stage when words with the most predictive power were identified. We use a probability threshold of 0.80 and classify any reports below this level as fraudulent. The last line of the table provides the area under the Receiver Operating Characteristic (ROC) curve as an alternative measure of how effectively the probability-of-truth measure is able to distinguish between fraudulent and truthful reports. A ROC area of 0.50 implies that a characteristic is no better than a random classifier while measures above this value indicate higher levels of classification power.

the sum of correctly identified fraudulent and truthful reports divided by the total number or reports, is in excess of 82 percent. Results are similar across both samples with a slightly higher proportion of truthful reports being correctly classified (83 percent) in comparison to fraudulent reports (80 percent). In both samples the area under the ROC curve is 0.89. To help put these numbers in perspective, recall that 77 percent of our sample represents truthful reports implying that a naïve classifier always predicting truth would correctly identify 77 percent of all observations in contrast to the 82 percent that our

measure classifies. More importantly than this improvement, however, is the ability of our approach to identify 80 percent of frauds. These frauds would have been completely missed by a naïve truthful classifier.

Our preliminary examination of classification rates suggests that the text of the MD&A section can provide important indicators of financial misrepresentation and flag suspicious instances that investors or auditors may wish to investigate in more detail. We turn now to providing a thorough comparison of this approach to other fraud-detection tools. Since many of these tools have not (or cannot) be extended to quarterly reports, we follow previous work and confine our comparison to the effectiveness of these tools in identifying fraudulent 10-K reports. We do this in two settings, first restricting the analysis to the MD&A sections of annual reports contained in our original training sample reflecting the truthful and fraudulent reports from a group of firms with at least one AAER bulletin issued against them. In our second set of comparisons, we change the sample to conform to a one-to-one matching approach in which each fraudulent annual report is matched to a report from a similar firm with no history of AAER bulletins. This type of matching is common in fraud-detection research as evidenced by its use by Erickson, Hanlon, and Maydew (2006), Brazel et al. (2009), Glancy and Yadav (2011), and many others.

4. Comparison of fraud detection methods within 10-K reports

Discussion of alternative quantitative methods

We compare the effectiveness of the probability-of-truth variable to four quantitative measures and four predefined language dictionaries previously suggested to be linked to financial misrepresentation. These measures and their sources are summarized in Table 2. The first measure, the *F*-score (Model 1) from Dechow et al. (2011) encompasses many financial statement variables that have previously been suggested to be associated with financial misstatements including accruals, financing activity, return on assets, the percentage of soft assets and changes in receivables, inventories, and cash sales.[13] Coefficients relating these variables to observed financial misrepresentations are estimated via a logit model and used to create the *F*-score which compares the expectation of fraud conditional on these firm characteristics to an unconditional expectation. The version of the *F*-score presented by Dechow et al. (2011) is based on AAER bulletins from 1982–2005. To be consistent with our time period, we calculate the coefficients for the *F*-score using all annual reports available in COMPUSTAT between the years 1994 and 2006, resulting in approximately 65,000 reports with sufficient information for calculating the *F*-score. Within our sample used to develop the probability-of-truth measure, we find that 802 or approximately 71 percent of annual reports have sufficient data for estimating the *F*-score and therefore these firms constitute the primary sample for comparing the alternative fraud-detection techniques.

Since the *F*-score was developed on the basis of an extensive evaluation of financial statement variables, we look for alternative, nonfinancial measures as possible indicators of fraud. For instance, our second measure, capacity differential, comes from Brazel et al.'s (2009) examination of nonfinancial measures such as facility growth which logically should ultimately be reflected in revenue growth. If revenue is seen to grow without a substantial increase in the facilities needed to produce this revenue, it may be an indication that managers are fraudulently overstating revenue. Our specific measure of capacity difference is the difference between the year-over-year change in revenue less the corresponding change in the number of employees.[14] In a logistic regression, Brazel et al.

13. We do not examine Models 2 and 3 of the *F*-score since the inclusion of the additional variables reduces the sample size while doing little to improve classification rates.
14. As in Brazel et al. (2009) we winsorize the capacity difference to lie between −100 percent and +100 percent.

TABLE 2
Summary of alternative fraud-detection techniques

Technique	Description	Source
Quantitative detection techniques		
F-score	Ratio of the predicted probability of fraud based on a logit model of firm financial characteristics, over the unconditional expectation of a financial misstatement.	Dechow et al. (2011)
Capacity difference	Difference between the year-over-year change in revenue less the year-over-year change in number of employees.	Brazel et al. (2009)
M&A activity	Dummy variable equal to one if the firm acquires a target or merges with another firm during the fiscal year. The variable is otherwise equal to zero.	Brazel et al. (2009)
Unexplained audit fees	The residual from a regression of characteristics previously shown to be associated with audit fees on the log of audit fees.	Hribar et al. (2010) with modifications proposed by Price et al. (2011)
Language-based detection techniques		
Deceptive proportion	Frequency of appearance in the MD&A section of words from a list of deceptive words corresponding to the categories of first person singular pronouns, exclusive words, negative-emotion words, and action verbs. The frequency is then divided by the total number of words in the MD&A.	Newman et al. (2003)
Litigious proportion	Frequency of appearance in the MD&A section of words from the Fin-Litigious list. The frequency is then divided by the total number of words in the MD&A.	Loughran and McDonald (2011)
Uncertain proportion	Frequency of appearance in the MD&A section of words from the Fin-Uncertain list. The frequency is then divided by the total number of words in the MD&A.	Loughran and McDonald (2011)
Negative proportion	Frequency of appearance in the MD&A section of words from the Fin-Negative list. The frequency is then divided by the total number of words in the MD&A.	Loughran and McDonald (2011)

(The table is continued on the next page.)

TABLE 2 (continued)

Technique	Description
Probability of truth	Measure of how likely a report is to be truthful based on classification by support vector machines (SVM). The presence in the MD&A of the top 200 words most predictive of fraud provide the inputs for the SVM classification. Predictive words are identified from a training sample and not known ex ante.

Notes:

The table provides a brief description of the nine alternative fraud detection techniques compared in this paper and their source. We divide these techniques into those based on financial/quantitative measures and those using textual analysis. For reference the probability-of-truth measure derived in this paper is also described.

(2009) find this measure to be positively associated with the likelihood of fraud. They find a similar association between M&A activity and fraud so we include this as an alternative predictor of fraud, collecting data on our firms' involvement in M&A transactions from the SDC Disclosure database. Using this information we create a dummy variable equal to one for years in which a firm was an acquirer or merged with another firm and zero otherwise. For simplicity we refer to the M&A dummy as a third quantitative indicator of fraud to distinguish it from language-based approaches.

Our final quantitative fraud indicator is unexplained audit fees as developed by Hribar et al. (2010). The motivation behind this measure is that auditors will charge higher fees for firms they suspect to have poor-quality earnings. These higher fees reflect compensation for the risk of litigation if, in fact, a bad accounting outcome occurs. The unexplained portion of fees represents the residual from a regression of characteristics previously known to be associated with fees on the log of audit fees.[15] We follow the slight modifications to the original Hribar et al. (2010) model suggested by Price et al. (2011) to ensure the greatest amount of data available but, even so, our sample size falls significantly for this measure given that it relies on the availability of audit fees from the Audit Analytics Database which has a starting date of 2000.

Discussion of alternative language-based methods

In addition to the four quantitative measures described above, we compare the performance of our data-generated list of predictive words to four fixed-word lists: the Newman, Pennebaker, Berry, and Richards (2003) deception model and three lists created by Loughran and McDonald (2011) to reflect tone in business documents which they found to have positive associations with financial misrepresentation (Fin-Negative, Fin-Uncertainty, and Fin-Litigious).

The deception model is based on 86 words in four categories: first-person singular pronouns; exclusive words (which signal refinements to content such as *but* and *or*);

15. The full list of characteristics that we include to explain audit fees is 2-digit SIC industry indicator, Big N accounting firm, *Log(Assets)*, a dummy variable equal to one if the firm experienced losses in either of the two previous years, a dummy variable equal to one if the auditor's opinion was anything other than unqualified, and inventory, receivables, long-term debt, and earnings all scaled by average assets.

negative-emotion words; and action verbs. The list was generated based on an analysis of text from individuals aware that they were providing deceptive information versus those providing truthful messages.[16] The word lists from Loughran and McDonald vary significantly in size with the Fin-Negative being the largest at over 2,300 words and the Fin-Uncertainty the smallest at 292 words. The lists are not mutually exclusive. For instance, the word "loss" is included in both the deceptive and Fin-Negative bags of words. Similarly, the word "litigation" appears on both the Fin-Negative and Fin-Litigious lists.

To get some sense of the various word lists and the presence of these words in our sample, Table 3 provides the ten most frequently occurring words from each of the four lists within the MD&A section of our annual and quarterly reports. In the second portion of the table we list words that overlap between each of the four fixed-word lists and our data-generated list of the 200 most predictive words. Although the word "loss" appears on the Fin-Negative, Deceptive, and data-generated word lists, the lists are generally quite distinct. At most the data-generated list produces three overlapping words with any of the

TABLE 3
Common and overlapping words

Deceptive Model (86 words)	Fin-Negative (2,350 words)	Fin-Uncertainty (292 words)	Fin-Litigious (871 words)
Panel A: Most frequent words from each word list within the sample			
or	loss	approximately	contracts
loss	restructuring	may	contract
however	losses	could	litigation
but	impairment	may	claims
although	adverse	must	settlement
carrying	decline	possible	legal
action	adversely	depend	regulatory
without	litigation	might	contractual
driven	claims	depends	laws
except	against	uncertain	court
Panel B: Overlapping words from fixed lists and data-generated list			
loss	critical	approximately	contract
or	loss	believes	legal
		may	settlement

Notes:

The table provides the 10 most frequent words occurring in our sample from four fixed-word lists previously linked to fraud. Note that multiple occurrences of the same word may be listed if the word is used in a variety of ways corresponding to different meanings. The first list refers to the deceptive model of Newman et al. (2003) while the remaining three lists are developed by Loughran and McDonald (2011) to refer to financial documents specifically. Panel B of the table provides words that appear on each of the four lists that are also present among the 200 words with the greatest predictive power as established by our decision tree approach. This approach makes no ex ante claim about the predictive ability of any word but allows the data to identify which words are associated with truth and fraud from a training sample including both truthful and fraudulent reports.

16. See Tausczik and Pennebaker (2010) for a review of the extensive use of this word list in a variety of contexts.

fixed-word lists suggesting that what the data identifies to be a predictive word may be difficult for researchers to identify ex ante.

Under the bag of words methodology, several alternative methods can be used to establish the significance of the appearance of words contained on a predefined list. A common and straightforward approach is to count the number of times that words on the list appear and divide this count by the total number of words to provide a proportionate measure (Bentley, Omer, and Sharp 2012; Davis et al. 2012). Alternatively, various weightings can be used to try to adjust for the presence of common versus infrequently used words (Loughran and McDonald 2011). In our comparisons, we used two approaches to establishing the merits of these word lists. First, we examined the associations between simple proportionate counts and reports' status as either truthful or fraudulent. Second, we examined how effective the various lists were as the primary inputs into step three of our method which generates a probability-of-truth measure for each report. As the insights gained from both approaches were similar, we present only the more straightforward proportionate word counts in our comparisons.

Association of the various methods with fraudulent MD&A sections in 10-K reports

Table 4 provides summary statistics and correlations among the nine fraud-detection measures compared across 10-K reports. Panel A of the table provides the mean and median values for the measures reported separately for truthful and fraudulent reports. The last two rows of the panel formally test for differences in these mean values and provide the ROC area assessing their overall ability to correctly differentiate truth from fraud. Panel B of the table provides the correlation among the nine methods.

The summary statistics show statistically significant differences in means for all of our measures except the M&A indicator dummy and Brazel et al.'s (2009) capacity differential measure. Despite their statistical significance, however, many of the measures do not behave as hypothesized. While the proportion of uncertain words is higher in fraudulent MD&A sections compared to truthful ones, we see higher average proportions of deceptive, litigious, and negative words in truthful rather than fraudulent reports. Perhaps this is due in part to our inclusion of annual reports for some firms in periods after a fraud has been identified. It is possible for instance that these reports would be required to disclose the negative and/or legal outcomes associated with the fraud whereas reports written during the time of the fraud may paint an overly optimistic impression of firm performance, thereby reducing the number of negative words. While a case can be made for either positive or negative associations between fraud and the various word lists, we see that the proportion of words used from any one of these lists represents a small fraction of words in the MD&A. The largest of the lists, Fin-Negative, represents approximately 1.5 percent of words used in fraudulent reports in comparison to 1.6 percent of truthful MD&A sections. The relatively low proportionate counts suggest that for some reports there may be no occurrence of words deemed suspicious by the various word lists. We examine whether this is the case and find it to be most problematic for the Fin-Litigious list. For annual report MD&As we find that approximately 2 percent of fraudulent reports do not contain a single word from this list. Extending this to interim reports we find that approximately 6 percent of fraudulent and 5 percent of truthful reports make no mention of these words suggesting that this particular word list may not be highly effective in distinguishing truth and fraud for at least a portion of reports.

Turning to our data-generated probabilities of truth, we continue to see a significant difference in the probability measures associated with truthful versus fraudulent reports (87.7 percent and 65.2 percent respectively) in this smaller sample restricted to 10-K reports alone. The ROC area reported is 0.87, just slightly below the 0.89 result for the entire sample of both 10-K and 10-Q reports. The *F*-score behaves as expected with

TABLE 4
Summary statistics and correlation

Panel A: Summary statistics for alternative prediction techniques across annual reports

	Probability of truth	*F*-score	Capacity difference	M&A activity	Unexplained audit fees	Deceptive words	Litigious words	Uncertain words	Negative words
Truthful									
Mean	0.877	1.237	0.066	0.017	0.478	1.190%	0.636%	1.314%	1.616%
Median	0.892	0.950	0.049	0.000	0.489	1.130%	0.588%	1.238%	1.551%
N	638	638	608	638	291	638	638	638	638
Fraudulent									
Mean	0.652	1.832	0.090	0.018	−0.208	1.095%	0.502%	1.451%	1.474%
Median	0.704	1.512	0.067	0.000	−0.181	1.041%	0.434%	1.273%	1.358%
N	164	164	157	164	61	164	164	164	164
T-stat for mean diff.	19.38**	−6.16**	−0.85	−0.09	6.06**	2.75**	4.30**	−2.38*	2.22*
(*p*-values)	(0.00)	(0.00)	(0.40)	(0.93)	(0.00)	(0.01)	(0.00)	(0.02)	(0.03)
ROC area	0.87	0.66	0.53	0.50	0.28	0.44	0.40	0.53	0.45

(The appendix is continued on the next page.)

Panel B: Pairwise correlation between alternative prediction methods across annual reports

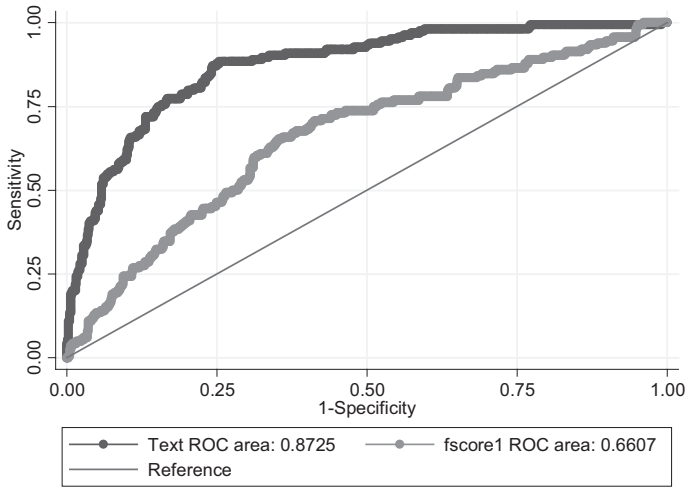
	Probability of truth	*F*-score	Capacity difference	M&A activity	Unexplained audit fees	Deceptive words	Litigious words	Uncertain words	Negative words
Probability of truth	−1.00								
F-score	−0.246**	1.00							
Capacity difference	−0.021	0.008	1.00						
M&A activity	−0.068	0.069	−0.012	1.00					
Unexplained audit	0.301**	−0.094	−0.055	−0.033	1.00				
Deceptive words	0.113**	−0.020	0.004	−0.061	0.033	1.00			
Litigious words	0.223**	−0.149**	0.051	−0.069	0.262**	0.094**	1.00		
Uncertain words	−0.091**	−0.007	0.028	−0.007	0.009	0.158**	0.164**	1.00	
Negative words	0.144**	−0.210**	−0.004	−0.018	0.192**	0.241**	0.471**	0.364**	1.00

Notes:

* Represents significance at the 5 percent level while ** represents significance at the 1 percent level.

The table provides summary statistics and pairwise correlations for the nine alternative fraud-detection methods we examine within a sample of 802 annual reports. Panel A provides the mean and median values of each measure for the subsamples of truthful and fraudulent financial reports. Reports are identified to be fraudulent on the bases of AAER bulletins issued by the SEC. *t*-tests for differences in the means of each fraud indicator between truthful and fraudulent reports are presented in addition to the area under the Receiver Operating Characteristic (ROC) curve. The alternative fraud detection measures are summarized in Table 2. The sample for the unexplained audit fee measure is significantly smaller due to the unavailability of the required data prior to the year 2000.

Figure 1 Area under the ROC curve: Probability of truth versus *F*-score



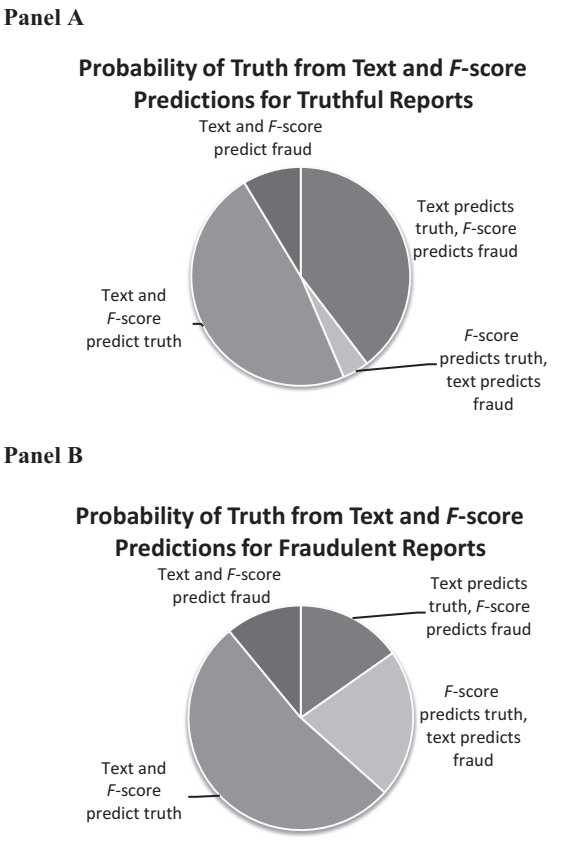
Description:
The figure plots the area under the Receiver-Operating-Characteristic (ROC) curve for our text-based probability-of-truth measure and Dechow et al.'s (2011) *F*-score (Model 1). The two fraud indicators are applied to a sample of 802 annual reports from firms with at least one AAER bulletin asserting fraud. ROC curves plot the true positive rate (sensitivity) versus the false-positive rate (1-Specificity). An area under the curve greater than 0.50 implies that the fraud indicator has some ability to correctly classify fraudulent and truthful reports with larger numbers corresponding to improved classification. A value of 0.50 implies that the indicator has no more power than a random classifier.

significantly higher values for fraudulent reports than truthful ones. The *F*-score produces the second highest ROC area at 0.66. For comparison to our data-generated probability-of-truth measure, we plot the two ROC curves in Figure 1.

The difference in unexplained audit fees is significant, however not in the direction that we would expect, and the ROC area indicates a poor degree of classification power for this measure. However, we should keep in mind that the sample size decreases significantly for this method due to the unavailability of the Audit Analytics database prior to the year 2000. The correlation table in panel B provides more interesting results for this measure. While it is negatively related to financial statement measures of fraud such as the *F*-score and capacity differentials, it is positively associated with many of the word counts from the various lists. These correlations imply that it is a quantitative measure that may bring alternative insights from what is generally found by other financial or nonfinancial measures and that these insights appear somewhat consistent with language-based methods. It remains to be seen in larger samples, however, whether these insights can effectively identify fraud.

The correlation table suggests that while the language-based tools show significant positive correlation among each other (with the exception of Fin-Uncertain which shows *negative* correlation) the quantitative measures of *F*-score, capacity difference, and M&A involvement are not strongly related to one another. The positive correlation between the probability-of-truth measure and the deceptive, litigious, and negative word list frequencies again demonstrates the difficulty in establishing the direction of association between fraud and the appearance of words ex ante since we had expected a positive association between the word list counts and fraudulent reporting.

Figure 2 **Panel A:** Truthful reports
Panel B: Fraudulent reports



Description:
The figure compares predictions from the Dechow et al. (2011) *F*-score and our probability-of-truth measure based on textual analysis of the MD&A section of 10-K reports. We consider an *F*-score greater than 1 and a probability-of-truth measure below 80 percent to predict fraud. Panel A of the figure examines the performance of both fraud-detection methods for10-K reports that are true while panel B shows predictions for false reports. The figures show where the two fraud detection methods agree in their predictions and where they differ. While there is significant overlap in their predictions, each is able to correctly classify some reports that the other misses. The probability-of-truth measure appears to be able to reduce the number of false-positive predictions made by the *F*-score. The reports are from a sample of 802 10-Ks between the years 1994 and 2006 issued by companies with at least one AAER bulletin asserting fraud.

The low level of correlation between the quantitative measures is somewhat by design as we avoided additional fraud indicators based on financial statement variables alone since many of these are incorporated within the *F*-score. The *F*-score itself is negatively and significantly correlated with the probability-of-truth measure as expected if both are indicative of fraud. Given that these two measures have the highest ROC area, it is interesting that their correlation is relatively low at −0.25. This suggests that there may be ample room for these two measures to serve as complementary fraud-detection techniques rather than substitutes.

We explore the degree of complementarity between the *F*-score and our probability-of-truth measure more deeply. Figure 2 plots the predictions of each approach separately

for truthful reports (Panel A) and fraudulent ones (Panel B). Each panel provides a pie chart indicating the proportion of times the two methods either agree or disagree on the prediction for a given observation and whether that prediction is in fact correct.

From panel A we see that almost 48 percent of the time that a report is truthful, both methods successfully identify it to be so. Looking to panel B, we see that agreement is even higher for misrepresentations, with over 52 percent of fraudulent observations being correctly identified by both methods. In contrast, when a report is in fact truthful, both methods falsely predict fraud only 8.62 percent of the time. This is interesting in two respects; first, by showing that a prediction of fraud by both methods gives a powerful signal that the report is most likely fraudulent, and second by showing some overlap in false-positive predictions as well. While clearly it is possible that both models err in making these false-positive predictions, it is also possible that they have identified suspicious reports worthy of additional investigation. Also of note is the significant difference in the number of false positives generated by the two methods. Looking at panel A representing truthful reports, we see that a significant portion of these reports are incorrectly classified by the *F*-score to be fraudulent while correctly identified as truthful by the data-generated language list. The relatively high proportion of false positives generated by the *F*-score in our sample is consistent with Dechow et al.'s (2011) Type I error rate of approximately 36 percent. While the *F*-score may provide an efficient first screen of annual reports to identify suspicious cases, it appears that text may be used to further refine the list so that investigative resources can be best allocated to those instances most likely to represent actual financial misstatements rather than improperly accusing truthful disclosures. A final observation from the figure is that only 11 percent of frauds go undetected by either method, indicating that textual analysis in combination with an evaluation of quantitative accounting figures can capture the vast majority of financial misrepresentations.

We next compare the association of our nine alternative fraud detection methods to actual incidents of fraud through a logit analysis in Table 5. Similar to Bentley et al. (2012), our dependent variable represents whether or not a report contained financial misstatements while our explanatory variables include our suggested predictors of these misstatements. We compare the effectiveness of the approaches across two samples with the left side of the table, continuing to focus on annual reports from our original sample of truthful and fraudulent 10-K reports. The right side, however, changes the sample to work with the fraudulent annual reports matched to annual reports from firms without a history of fraudulent reporting as in Erickson et al. (2006), Brazel et al. (2009), and many others. Since our research seeks to establish whether language analysis can provide valuable incremental information to fraud investigators, we match on characteristics that would traditionally make it difficult to distinguish fraudulent reports from truthful ones. Given the strength of the *F*-score in our analysis, we choose a matching truthful report from the same industry and year with the closest available *F*-score from all firms in the COMPUSTAT database for each of the 164 fraudulent 10-Ks from our sample with sufficient data to calculate the *F*-score. Constructing matched pairs in this way allows us to see whether textual analysis or nonfinancial measures can be helpful in situations where traditional accounting based tools will struggle.

Panel A of Table 5 presents four alternative logit model specifications. The first two include the quantitative measures of fraud detection in addition to our data-generated language tool. The sole difference between the two columns is that the second includes the unexplained audit fee measure while the first does not. We present these results separately due to the much smaller time period for which the unexplained audit fee is available. The third specification in the table compares the association of the five language-based measures to fraudulent reports, while the fourth includes all measures, both quantitative and language-based, that are available for the entire sample period. In all models standard

TABLE 5
Logit analysis of alternative prediction methods

	Panel A: 802 annual reports (truthful and fraudulent) from firms named in AAER bulletins				**Panel B:** 164 fraudulent annual reports and 164 matches from firms never named in AAER bulletins			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Probability of truth	−8.63** (0.00)	−12.52** (0.00)	−8.49** (0.00)	−8.58** (0.00)	−8.13** (0.00)	−9.08** (0.00)	−7.75** (0.00)	−8.40** (0.00)
F-score	0.20* (0.01)	0.05 (0.84)		0.23** (0.00)				
Capacity difference	0.31 (0.48)	0.64 (0.49)		0.32 (0.45)	0.89* (0.03)	0.73 (0.20)		0.83 (0.05)
M&A activity	−1.23 (0.41)	0.15 (0.96)		−1.16 (0.40)	−2.21* (0.01)	−1.02 (0.34)		−2.28* (0.01)
Unexplained audit		−0.53 (0.01)				0.61 (0.06)		
Deceptive proportion			−33.73 (0.28)	−32.68 (0.38)			−65.32 (0.13)	−81.74 (0.07)
Litigious proportion			−16.80 (0.53)	−24.52 (0.41)			82.37 (0.09)	79.78 (0.10)
Uncertain proportion			47.95* (0.01)	47.48* (0.02)			30.05 (0.21)	29.64 (0.26)
Negative proportion			6.30 (0.70)	17.03 (0.35)			27.49 (0.26)	30.49 (0.23)
Constant	5.21** (0.00)	8.71** (0.00)	5.13** (0.00)	4.71** (0.00)	6.35** (0.00)	7.33** (0.00)	5.47** (0.00)	6.22** (0.00)

(The table is continued on the next page.)

TABLE 5 (continued)

	Panel A: 802 annual reports (truthful and fraudulent) from firms named in AAER bulletins				**Panel B:** 164 fraudulent annual reports and 164 matches from firms never named in AAER bulletins			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Observations	765	342	802	765	313	122	328	313
Pseudo R^2	0.30	0.39	0.29	0.31	0.27	0.29	0.24	0.29
ROC area	0.87	0.92	0.86	0.87	0.83	0.86	0.81	0.84

Notes:

* Represents significance at the 5 percent level while ** represents significance at the 1 percent level.

The table presents coefficient estimates (and *p*-values) from a logit model in which the dependent variable is equal to one if a report is classified as fraudulent by an AAER bulletin and zero otherwise. The independent variables represent alternative fraud-detection measures as described in Table 2. Panel A of the table conducts the logit model on a sample of 802 annual reports from the years 1994 and 2006 from firms with at least one fraudulent report identified by in an AAER bulletin. Standard errors are adjusted for clustering due to the presence of multiple observations from the same firm. Panel B of the table reduces the sample to the 164 fraudulent 10-K reports from the original sample matched to annual reports from firms never named in AAER bulletins. Matches are chosen on the basis of 2-digit SIC codes, year of observation and the closest possible Dechow et al. (2011) *F*-score from all possibilities in the COMPUSTAT database. Model 2 is based on the shorter time horizon of 2000 to 2006 due to the availability of audit fee data.

errors are adjusted for clustering to account for multiple observations form the same firm. We again report the ROC area as a measure of classification success; however, it should be noted that here it is associated with the success of the combination of alternative techniques presented in each model specification rather than any individual one.

The first observation from the table is that in all model specifications the probability of truth shows a strong negative association with actual fraudulent reporting, maintaining this association despite the inclusion of our eight alternative measures. Among the quantitative measures examined in panel A, we see again that the *F*-score emerges as the strongest alternative with positive and significant associations with fraud in all specifications except the smaller sample presented in column 2. Within this specification, the unexplained audit fee variable shows a significant association with fraudulent reporting but not in the direction we would expect.

Among the fixed-word lists, we see from model specifications three and four in panel A that the proportion of words from the Fin-Uncertain list is the only one to show a positive and statistically significant association with fraudulent reports. Consistent with the summary statistics presented in Table 4, a higher proportion of uncertain words appears indicative of fraudulent reporting while the same cannot be said of the litigious, negative, or deceptive word lists. This is particularly interesting in the case of deceptive words as it suggests that either deception in business documents is conveyed differently due to differences in the style of language used in these formal reports, or that at least some of the individuals involved in drafting the report were unaware that fraudulent activity was occurring and therefore the text exhibits no signs of conscious deception.

Moving to the matched sample comparison reported in panel B provides some interesting results. Here the *F*-score is dropped from the analysis given that it is used as the basis for finding truthful matches for each fraudulent 10-K so we do not expect it to be able to distinguish deceptive reports. In this context, we again see the strength of the data-generated word list through its strong negative association between the probability of truth and fraudulent reporting.[17] The logit models in panel B confirm the relatively poor performance of the fixed-word lists in correctly distinguishing between fraud and truth. While the proportion of uncertain words showed some promise for annual reports issued by firms with at least one AAER bulletin in panel A, it struggles to do so in a sample containing reports from firms without a history of problem reporting. The Fin-Negative, Fin-Litigious, and Deceptive word lists also show no clear association with fraud in this smaller sample, although their performance improves somewhat in that both the deceptive and litigious word lists now hover near the 10 percent significance level. While the deceptive words continue to show a higher proportionate count for truthful reports, the litigious word count now behaves as expected with a higher proportion of litigious words being positively associated with fraudulent reports. This suggests that these lists may perform better in a broad cross-sectional analysis identifying fraudulent firms from truthful ones but may struggle to pinpoint the exact timing of fraud in a time-series analysis of reports from a single firm.

The capacity difference measure also shows improvement in identifying fraud from truth in the context of a matched sample. Now, it provides a clear positive association with instances of fraud in all specifications except the very small sample including the unexplained audit fee measure. For specification two, the unexplained audit fees now

17. The overall classification rate in the matched sample is 71.04 percent in comparison to 82.19 percent achieved in the training sample. This rate exceeds that achieved by Dechow et al. (2011), Humphreys et al. (2011), and seven of Cecchini et al.'s (2010) eleven proposed models using single and multi-word phrases to identify fraud.

behave consistently with the work of Price et al. (2011) and demonstrate a positive association with fraud again suggesting that some fraud detection tools may be more effective in the cross-section than time series. Only the data-generated probability-of-truth measure shows a similar degree of power in both settings. As a result, we turn now to examine in more detail its performance across a time-series sample of both interim and annual reports.

5. Assessing language-based tools in time series

While Dechow et al. (2011) and Kaminski et al. (2004) show that a firm's financial variables show few systematic differences between times of fraud and the periods surrounding them, we turn to language-based tools to establish whether they can be effective in identifying fraud within a time-series sample of reports with a particular emphasis on our probability-of-truth measure. As a first step, we examine how the probability of truth changes over time for a firm involved in fraudulent financial reporting. Figure 3, panel A plots the mean and median probability of truth for the four quarters before and after an instance of fraud. We label fraudulent reports as occurring at time $T = 0$ acknowledging that this may in fact include multiple reports if misreporting occurs in consecutive 10-Q and 10-K reports. Time $T - 1$ represents the first quarter prior to a fraud involving either a single or multiple quarters while time $T + 1$ represents the first truthful report following a single or the last of a series of financial misrepresentations.

From Figure 3, we see that the probability of truth hovers around the overall average value of 80 percent until about two quarters prior to the first fraudulent report. At time $T - 2$ and $T - 1$, the probability falls below this value but not dramatically so, with median values remaining above 75 percent in both periods. The most dramatic plunge in this probability occurs from time $T - 1$ to time 0 as the median value plummets to 57 percent. The probability quickly recovers to normal levels following the fraud and returns to a median value of 81 percent by time $T + 1$.

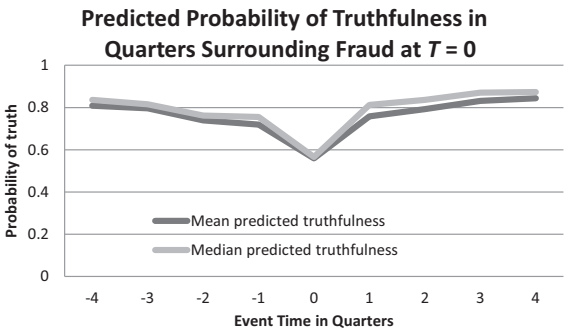
The relatively quick and dramatic dip in the probability of truth during times of fraud is consistent with the lower false-positive rates associated with this approach. Panel B of the figure plots precisely where the false positives fall in event time surrounding the fraud. Recalling from Table 1 that the false-positive rate for the entire sample was 17.22 percent, we see from panel B that these are clustered in the four quarters immediately prior to and following fraud, peaking at $T - 1$, the quarter prior to the first instance of fraud. In total, over 55 percent of all false positives occur within plus or minus a year of the fraud; however, these are more heavily weighted towards false positives in the preceding period with the four quarters leading up to the event responsible for over 31 percent of all false-positive findings.

Several interpretations of these results are possible. Most pessimistic, from the perspective of a fraud investigator, is that while our approach shows an ability to identify fraud in both quarterly and annual reports, language, like quantitative measures, struggles to correctly identify the precise timing of fraud. It could be that the similarity in a firm's MD&A discussion from one period to the next makes it difficult to pinpoint the moment that fraud begins. A more optimistic interpretation is that the gradual decline in truthfulness probability leading up to the fraud suggests that the cut-and-paste nature of financial reporting may result in a noticeable transition from exaggeration to misrepresentation. If this is the case, strings of successive false positives may be useful in the early identification or even prevention of misrepresentation. A final possibility is that our results have flagged frauds that the SEC is either unaware of or unable to prosecute effectively.

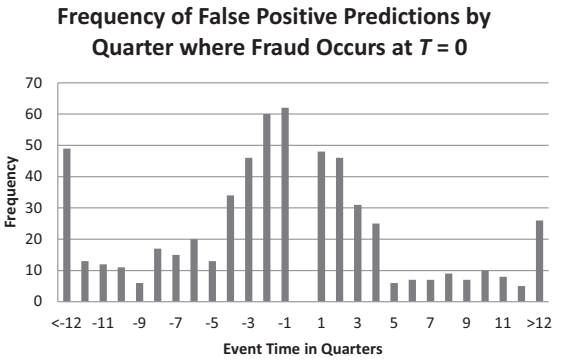
We explore in more detail the possibility that the decline or shift in the probability of truth is an important indicator of fraud and provides incremental power to fraud investigators. Unrelated to fraud, Feldman et al. (2010) demonstrate that the change in tone of a

Figure 3 **Panel A:** Probability of truth in event time
Panel B: False-positive observations surrounding fraud

Panel A



Panel B



Description:
Panel A of the figure shows the mean and median probability-of-truth measure as established from the text of the MD&A sections of 10-K and 10-Q reports. Time 0 represents the occurrence of the fraud, which may involve a single or multiple quarters. Time $T + 1$ is the first quarter following fraudulent reporting. We see that the probability-of-truth measure has a sudden and dramatic drop from the quarter preceeding a fraud to the first fraudulent quarter. Panel B of the figure plots the frequency of false-positive predictions generated by the probability-of-truth measure, that is, predictions that a report is fraudulent when in fact it is true. Again treating $T = 0$ as the occurrence of fraud, we see that the majority of false-positive predictions fall within one to four quarters surrounding the fraudulent incident.

firm's 10-Q and 10-K reports from previous filings helps explain stock returns even after controlling for financial information. Similarly, given our pattern for the probability of truth for quarters surrounding a fraud, we suggest that the change in this probability may be an important indicator of misrepresentations.

Table 6 explores this possibility and the overall effectiveness of language-based tools in a time-series sample of both quarterly and annual reports. We again include our probability-of-truth measure in addition to the proportion of words from the deceptive, negative, uncertain, and litigious word lists in a logit model with the dependent variable identifying fraudulent versus truthful reports. As in our previous samples based on annual reports alone and matched truthful firms, the first column of Table 6 shows little benefit for proportion word counts in identifying fraud. None of the four word lists show a

TABLE 6
Logit analysis of alternative text predictors in interim and annual reports

	(1) Coefficients (*p*-value)	(2) Coefficients (*p*-value)	(3) Odds ratio
Probability of truth	−9.35*	−12.74*	
	(0.00)	(0.00)	
Change in probability		6.16*	1.06
		(0.00)	
1-Probability of truth			1.14
Deceptive proportion	5.30	11.94	1.13
	(0.62)	(0.36)	
Litigious proportion	−37.13	−28.83	0.75
	(0.07)	(0.24)	
Uncertain proportion	8.71	9.32	1.10
	(0.37)	(0.43)	
Negative proportion	6.67	0.20	1.00
	(0.50)	(0.99)	
Constant	5.77*	8.30*	
	(0.00)	(0.00)	
Observations	4,895	4,449	4,449
Pseudo R^2	0.39	0.50	0.50
ROC area	0.89	0.93	

Notes:

* Represents significance at the 1 percent level.

The table provides the estimates from a logit model in which the dependent variable is equal to one for financial reports labelled fraudulent by an AAER bulletin and zero otherwise. The sample includes 4,895 annual and quarterly reports over the years 1994–2006 from a sample of firms each of which has at least one fraudulent report. The independent variables are potential indicators of fraud representing our probability-of-truth measure, the change in this probability of truth from one quarter to the next for a given firm, and the proportion of words appearing in the MD&A section from fixed-word lists indicative of deception, litigation, uncertainty, or negative sentiment (as described in Table 2). The first two columns of the table present the coefficient estimates (*p*-values) measuring the association between fraudulent reports and each potential indicator of fraud. The third column presents the odds ratio for each dependent variable from Model 2. To ease interpretation of the odds ratio we alter the probability-of-truth measure to be 1-probability of truth (or essentially, the probability of fraud). We see that keeping everything else fixed, a one percentage point increase in this value increases the odds of the observation being fraudulent by 14 percent. The change in probability of truth from one quarter to the next provides incremental predictive power and further enhances our ability to distinguish between fraudulent and truthful financial reports.

strong statistical association with fraud even in this larger sample although the probability-of-truth measure maintains its strong significance.

The second column of the table includes the same five language-based methods in addition to a measure of the change in the probability of truth. This change measure represents the difference in the probability of truth between two successive reports separated by a single quarter regardless of whether these are two quarterly reports or a quarterly and annual report. Given the significant and rather sudden drop in the probability

measures from time $T - 1$ to time $T = 0$, we expect a large positive change in probability (indicating that $\text{Probability}_{T-1} - \text{Probability}_{T=0} > 0$) to be strongly associated with our fraud indicator. Column 2 of Table 6 shows this to be the case with the change measure being strongly associated with fraud. This positive association does not reduce the influence of the level measure, as the probability of truth itself remains a statistically significant indicator of fraud in its own right.[18] Including the change measure improves both the pseudo R^2 and ROC area measures associated with the logit model including all language-based indicators but does little to alter our conclusions on the performance of the bag of word methods. We continue to find no significant association between fraud and the proportionate appearance of words from the deceptive, uncertain, negative, and litigious lists.

As a final indication of the incremental strength that the change in probability-of-truth measure can bring, we examine the same logistic regression from column 2 of the table but present the odds ratio associated with each variable as opposed to the coefficient estimates in column 3. To make the interpretation of the odds ratios more straightforward, we make a few simple adjustments to the variables. First we multiply all measures by 100 so that a unit change in the variable corresponds to a 1 percentage point change. For instance, the mean probability-of-truth measure transforms from 80 percent (or 0.80) to 80 so that the odds ratio can be interpreted as measuring the effect of a 1 percentage point change in this measure. In addition, we recast our probability-of-truth measure by measuring the probability of fraud or 1-probability of truth. By doing so, we ensure that the odds ratio for this variable exceeds 1 and illustrates the impact of a 1 percentage point increase in the probability-of-fraud on the likeliness of fraud occurring.

From the odds ratios in column 3, we see the influence of our two significant predictors, the probability-of-truth level and the change in this value from one period to the next. Focusing first on the level we see that if the probability of fraud increases by a single percentage point, the odds of the observation actually being fraudulent increase by 14 percent. Keeping the probability of fraud and all other variables fixed, an increase of 1 percentage point in the size of the change in the probability from one period to the next increases the odds of the observation being fraudulent by 6 percent.[19] That is, even if the probability-of-fraud measure were the same across two similar firms, knowing the path taken to arrive at this level and how large an incremental step it was from previous levels contains important information in distinguishing fraud from truth.

The time-series results presented in this section show that the statistical approach to textual analysis can be an effective tool for identifying fraud in a quarterly and time-series context. This is an important finding since some quantitative measures may not be available on a quarterly basis or may show insufficient variation from one period to the next. Researchers have either neglected interim reports entirely or claimed that they are unable to provide enough insight to aid in fraud detection. In contrast, our results show that a time series of these same reports may provide a useful benchmarking feature so that deviations in a firm's reporting style provide incremental predictive power in identifying fraudulent reports.

18. The correlation between the change and level measures of probability of truth is significant at 0.35. As a result, we repeat our logit analysis as an ordinary least squares regression in order to calculate the variance inflation factors (VIFs) and establish the extent to which multicollinearity between these two variables may be a problem. We find that all individual VIFs are below 2 with a mean value across all variables of 1.3. As a result, we conclude that multicollinearity does not influence our interpretation of the results.

19. It may be helpful to recall that, within this sample, the unconditional probability of an observation being fraudulent as measured by the number of fraudulent observations over all observations is 23 percent. This implies that the odds of an observation being fraudulent are 0.23/(1–0.23) or approximately 0.30.

6. Conclusion

We develop a language-based method for detecting fraud using the words in the MD&A sections of annual and interim reports. Since we make no ex ante judgements about what may or may not be an indicator of fraud, we train the model using AAER bulletins issued by the SEC, taking their distinctions between fraud and truth to identify financial reports with and without fraudulent misrepresentations. The extent to which the SEC misses instances of fraud or is biased in its investigative choices to focus on firms of a particular size or from a given industry, will reduce the applicability of our results. However, this limitation is shared by all papers relying on the AAER bulletins to identify fraud.

Beginning with all words used in our sample's MD&A sections, we rank-order their effectiveness as distinguishers of fraud from truth. We then use the top 200 most-predictive words and SVMs to classify reports as fraudulent or truthful and achieve correct classification rates as high as 82 percent.

We compare the effectivness of our method to alternative fraud detection approaches across different samples and find that it consistently performs well. Among the eight alternatives we examine, we find that the *F*-score from Dechow et al. (2011) is the next most effective indicator of fraud. Interestingly, we find that the probabilty of truth and *F*-score measures serve as complements to one another in that each method is able to identify some frauds that the other misses and their correlation is relatively low. The probability-of-truth measure can be particularly useful in reducing the large number of false positives generated by the *F*-score measure. Among the other alternatives, we find little evidence of effective fraud prediction for proportionate word counts of deceptive, negative, uncertain, or litigious lists. Other quantitative measures such as capacity differential and unexplained audit fees perform better in the cross-sectional sample of matched truthful and fraudulent firms than they do in the sample of truthful and fraudulent reports from the same set of firms. In other words, financial measures may be useful to help identify suspicious sets of firms but then struggle to establish the precise timing of any fraudulent behaviour within this set.

Given the similarity between financial statement numbers from one period to the next and their poor performance in distinguishing truthful from fraudulent reports for a given firm, we examine our language-based tools within a time-series sample of 10-K and 10-Q reports to establish whether language can be an effective identifier of fraud within interim reports. We again find strong support for our probability-of-truth measure. Moreover, in tracking this probability over time we see a slight decline in the two quarters preceeding the fraud; however, there is a dramatic dip in its value from $T - 1$ to $T = 0$ when a fraud occurs. We find that the change in the probability of truth can provide incremental power in indentifying fraud beyond our level measure alone suggesting that, much like accounting numbers, language from a firm's previous reports may serve an important benchmarking function and deviations from this benchmark may be an indicator of fraud. Our results suggest that future accounting research should continue to move from focusing on cross-sectional tests to incorporating additional temporal aspects of a firm's financial reporting and that textual analysis may be particuarly well-suited to this context.

We make several important contributions to the literature: generating a data-derived language tool that appears to be an effective predictor of fraud, conducting a thorough comparison across both quantitative and language-based detection methods, and providing the first indication that the inclusion of quarterly reports and a temporal measure of deviations from previous language used by companies may provide incremental power in identifying fraudulent reports in time series. These insights will help both academics seeking to extend work in the areas of textual analysis and fraud detection and practitioners looking to make informed choices on the detection methods they wish to implement.

Appendix 1

Technical procedure

In this appendix we provide technical details on achieving the three steps required for our detection algorithm.

1. *Create the word-frequency matrix*: We identify words based not only on their spelling but also on their part-of-speech usage. The QTagger tool developed by Creasor and Skillicorn (2012) applies tag fields from the Penn Treebank Project (http://www.cis.upenn.edu/~treebank/) where N represents words used as nouns, V verbs, JJ adjectives etc. QTagger relies on Lingpipe (http://alias-i.com/lingpipe) to correctly classify words based on their use in a particular context. We select the 1,100 most common words to be used in step two of the analysis. We experiment with various other subsets of the words: the 5,000 most frequent or the entire set of 22,392 unique words. We also experimented with eliminating or retaining the top 100 most frequent words as we hypothesized they were too common to provide any meaningful predictive power. We found that the addition of these words did little to alter our results, leading us to retain the top 100 most common words and the next 1,000.
2. *Produce the rank-order of word's predictive ability*: 25 percent of documents were excluded at random from the sample during this stage. Among the remaining documents, a random forest of 3,000 trees was created. For each tree, a bootstrapped sample from the remaining 75 percent of documents was chosen randomly with replacement. At each node of the tree, 55 words were selected for consideration to be used in the inequality at that node. When a particular document is not used in the building of a particular tree, it is classified as "out of bag" and the tree can be used to predict whether that report is fraudulent or truthful. The "out-of-bag" predictive accuracy of the Random Forest in this context is approximately 86 percent.

   Most software packages implementing the Random Forests algorithm can produce a list of variables rank-ordered according to their ability to correctly classify observations. The approach used here relies on randomly changing the order of values in a particular column of the word-frequency matrix. The intuition is this: if the values in a column (i.e., the frequencies of one word in all reports) are permuted randomly, and the forest makes predictions using this altered data, then accuracy should be much lower than the accuracy using the unaltered data if the word is in fact an important predictor. On the other hand, if the word is not very predictive, not much difference in the two predictive accuracies should be observed.

   While Random Forests can be used to directly to predict truth and fraud, we used support vector machines (SVM) as our predictive methodology both to avoid any systematic bias introduced by using Random Forests in two different roles and because SVMs have previously been shown to correctly classify text (Joachims 1998; Antweiler and Frank 2004). Burges' (1998) tutorial on SVM provides details on the numerous contexts in which the method has been applied, in addition to its strengths and weaknesses compared to alternative methods.
3. *Use support vector machines to classify reports*: Support vector machines use two important ideas to build good predictors. First, they construct a hyperplane separating the two classes (fraud and truth) whose position is determined by maximizing the thickness of a rectangular region between the classes and placing the hyperplane at the midline. Second, since such a hyperplane will not, in general, exist in the raw data, new combination attributes are built from the existing ones, increasing the apparent dimensionality of the space until a linear separation becomes possible. The Lib-SVM algorithm[20] was

20. See Chang and Lin (2011) for details.

used in Matlab with a radial basis function kernel, C = 70 and g = 0.01. These values were chosen after a parameter sweep, but the prediction performance is not strongly dependent on these exact values; for example, performance decreases slightly (a percentage point or so) if C is as small as 20 or as large as 100.

Appendix 2

Rank-ordered list of most predictive words

The words below were generated from the application of the Random Forests data mining technique (Breiman 2001) to a sample of truthful and fraudulent reports identified by the SEC's AAER enforcements. Words are rank-ordered from most to least predictive although their relationship with fraudulent or truthful reports may be more complicated than frequency alone. In other words, a high-ranking word does not necessarily imply that more frequent occurrences of it are associated with fraud or truth. In some instances the same word may appear more than once, indicating that the word exists as different parts of speech; for instance, the same word may be used both as a noun and as an adjective. Symbols also appear: for example, the $ symbol appears twice, first associated with a singular noun (e.g., $100 is a lot of money) and the second time associated with a plural noun (e.g., $100 were paid to bribe the auditor). For simplicity we do not identify how each word is used; however, in most cases the most common usage of the word applies. The letter *s* also appears and is associated with the use of the possessive *'s*. The apostrophe itself does not appear since it is indistinguishable from single and double quotation marks.

acquisitions	taxes	from
acquisition	capital	facility
months	revenue	business
sales	under	cash
legal	expenses	income
revenues	%	decrease
shares	credit	products
approximately	by	ended
settlement	profit	debt
agreement	price	with
operating	rate	activities
quarter	letters	acquired
sale	software	during
increased	contract	an
working	operations	have
expenditures	or	these
at	on	new
as	prices	businesses
company	development	is
customers	estimate	increase
decreased	for	its
fiscal	inventory	costs
s	may	through
net	existing	result
it	gross	current
no	compared	management
valuation	future	changes

(The appendix is continued on the next page.)

Appendix 2 (continued)

line	unit	administrative
receivables	services	environmental
in	results	time
other	were	percentage
product	significant	equipment
company's	average	cost
year	been	per
interest	was	december
total	expense	stock
first	margin	over
marketing	provided	same
factors	due	information
and	rates	loss
primarily	tax	deferred
market	will	accounts
be	number	$
of	$	a
payments	because	distribution
also	&	estimates
annual	lease	related
the	to	percent
there	assets	maintenance
additional	series	devices
are	financing	amount
general	federal	of
commissions	expected	demand
consolidated	net	benefit
end	facilities	domestic
any	their	investments
that	research	primary
including	used	critical
selling	up	programs
fees	trading	section
which	receivable	retail
liquidity	while	judgments
compensation	matters	reduction
companies	volume	due
period	believes	likely
revolving	financial	
management's	prior	
allowance		

References

Abbassi, A., C. Albrecht, A. Vance, and J. Hansen. 2012. MetaFraud: A meta-learning framework for detecting financial fraud. *MIS Quarterly* 36 (4): 1293–327.

Antweiler, W., and M. Frank. 2004. Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance* 59 (3): 1259–94.

Beneish, M. 1997. Detecting GAAP violations: Implications for assessing earnings management among firms with extreme financial performance. *Journal of Accounting and Public Policy* 16 (3): 271–309.

Beneish, M. 1999. The detection of earnings manipulation. *Financial Analysts Journal* 55 (5): 24–36.

Bentley, K., T. Omer, and N. Sharp. 2012. Business strategy, financial reporting irregularities and audit effort. *Contemporary Accounting Research* 30 (2): 780–817.

Brazel, J., K. Jones, and M. Zimbelman. 2009. Using nonfinancial measures to assess fraud risk. *Journal of Accounting Research* 47 (5): 1135–66.

Breiman, L. 2001. Random forests. *Machine Learning* 45 (1): 5–32.

Brown, S., and J. Tucker. 2011. Large-sample evidence on firms' year-over-year MD&A modifications. *Journal of Accounting Research* 49 (2): 309–46.

Burges, C. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2 (2): 121–67.

Canadian Public Accountability Board. 2012. Public Report. Available online at: http://www.cpab-ccrc.ca/EN/content/CPAB_Public_Report_2012_Eng.pdf, retrieved May 30, 2013.

Cecchini, M., H. Aytug, G. Koehler, and P. Pathak. 2010. Making words work: Using financial text as a predictor of financial events. *Decision Support Systems* 50 (1): 164–75.

Chang, C., and C. Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2 (3): 1–27.

Cole, C., and C. Jones. 2005. Management discussion and analysis: A review and implication for future research. *Journal of Accounting Literature* 24 (1): 135–75.

Creasor, J., and D. Skillicorn. 2012. QTagger: Extracting word usage from large corpora. Technical Report, Queen's University, School of Computing.

Davis, A., J. Piger, and L. Sedor. 2012. Beyond the numbers: Measuring the information content of earnings press release language. *Contemporary Accounting Research* 29 (3): 845–68.

Dechow, P., W. Ge, C. Larson, and R. Sloan. 2011. Predicting material accounting misstatements. *Contemporary Accounting Research* 28 (1): 17–82.

Donahue, B. 2013. How to fail at corporate fraud. *Threat Post: The Kaspersky Lab Security News Service*. Available at threatpost.com/en_us/blogs/how-fail-corporate-fratud-010813, accessed January 9, 2013.

Eaglesham, J. 2013. Accounting fraud targeted: With crisis-related enforcement ebbing, SEC is turning back to Main Street. *Wall Street Journal*, May 28.

Erickson, M., M. Hanlon, and E. Maydew. 2006. Is there a link between executive equity incentives and accounting fraud? *Journal of Accounting Research* 44 (1): 113–44.

Ernst & Young. 2010. Driving ethical growth—New markets, new challenges. 11th Global Fraud Survey. Available online at http://www.ey.com/Publication/vwLUAssets/Driving_ethical_growth__new_markets,_new_challenges:_11th_Global_Fraud_Survey/$FILE/EY_11th_Global_Fraud_Survey.pdf, retrieved June 18, 2013.

Feldman, R., S. Govindaraj, J. Livnat, and B. Segal. 2010. Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies* 15 (4): 915–53.

Glancy, F., and S. Yadav. 2011. A computational model for financial reporting fraud detection. *Decision Support Systems* 50 (3): 595–601.

Goel, S., and J. Gangolly. 2012. Beyond the numbers: Mining the annual reports for hidden cues indicative of financial statement fraud. *Intelligent Systems in Accounting, Finance and Management* 19 (2): 75–89.

Goel, S., J. Gangolly, S. Faerman, and O. Uzuner. 2010. Can linguistic predictors detect fraudulent financial filings? *Journal of Emerging Technologies in Accounting* 7 (1): 25–46.

Hanley, K., and G. Hoberg. 2012. Litigation risk, strategic disclosure and the underpricing of initial public offereings. *Journal of Financial Economics* 103 (2): 235–54.

Hribar, P., T. Kravet, and R. Wilson, 2010. A new measure of accounting quality. Available online at http://ssrn.com/abstract = 1283946 or http://dx.doi.org/10.2139/ssrn.1283946, retrieved June 18, 2013.

Humphreys, S., K. Moffit, M. Burns, J. Burgoon, and W. Felix. 2011. Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems* 50 (3): 585–94.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. Proceedings of the European Conference on Machine Learning.

Kaminski, K., T. Wetzel, and L. Guan. 2004. Can financial ratios detect fraudulent financial reporting? *Managerial Auditing Journal* 19 (1): 15–28.

Larcker, D., and A. Zakolyukina. 2012. Detecting deceptive discussions in conference calls. *Journal of Accounting Research* 50 (2): 495–540.

Li, F. 2008. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics* 45 (2–3): 221–47.

Li, F. 2010a. The information content of forward-looking statements in corporate filings: A naïve Bayesian machine learning approach. *Journal of Accounting Research* 48 (5): 1049–102.

Li, F. 2010b. Textual analysis of corporate disclosures: A survey of the literature. *Journal of Accounting Literature* 29 (1): 143–65.

Loughran, T., and B. McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance* 66 (1): 35–65.

Meiers, D. 2006. The MD&A challenge. *Journal of Accountancy* 201 (1): 59–66.

Newman, M., J. Pennebaker, D. Berry, and J. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin* 29 (5): 665–75.

Pennebaker, J. W., C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth. 2007. The development and psychometric properties of LIWC2007. Software manual. University of Texas at Austin and University of Auckland.

Price, R. III., N. Sharp, and D. Wood. 2011. Detecting and predicting accounting irregularities: A comparison of commercial and academic risk measures. *Accounting Horizons* 25 (4): 755–80.

Tausczik, Y., and J. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29 (1): 24–54.

Zhou, W., and G. Kapoor. 2011. Detecting evolutionary financial statement fraud. *Decision Support Systems* 50 (3): 570–75.