JOURNAL OF EMERGING TECHNOLOGIES IN ACCOUNTING
Vol. 7
2010
pp. 25–46

American Accounting Association
DOI: 10.2308/jeta.2010.7.1.25

Can Linguistic Predictors Detect Fraudulent Financial Filings?

Sunita Goel
Siena College

Jagdish Gangolly
Sue R. Faerman
Ozlem Uzuner
University at Albany, SUNY

ABSTRACT: Extensive research has been done on the analytical and empirical examination of financial data in annual reports to detect fraud; however, there is scant research on the analysis of text in annual reports to detect fraud. The basic premise of this research is that there are clues hidden in the text that can be detected to determine the likelihood of fraud. In this research, we examine both the verbal content and the presentation style of the qualitative portion of the annual reports using natural language processing tools and explore linguistic features that distinguish fraudulent annual reports from nonfraudulent annual reports. Our results indicate that employment of linguistic features is an effective means for detecting fraud. We were able to improve the prediction accuracy of our fraud detection model from initial baseline results of 56.75 percent accuracy, using a "bag of words" approach, to 89.51 percent accuracy when we incorporated linguistically motivated features inspired by our informed reasoning and domain knowledge.

Keywords: fraud detection; textual content; qualitative analysis; linguistic features.

INTRODUCTION

The string of corporate governance failures and accounting scandals that have occurred since 2001 has brought unprecedented attention to the importance of deterring fraud and its timely detection when it does occur. In an attempt to restore investor confidence and order in the financial markets, following the multi-billion dollar accounting failures at Enron and WorldCom, the U.S. Congress passed the Public Company Accounting Reform and Investor Protection Act of 2002, otherwise known as the Sarbanes-Oxley Act (SOX). Even though the congressional intent underlying SOX was investor protection, numerous cases of fraudulent financial reporting have surfaced since its passing (e.g., the AIG scandal, *BusinessWeek* 2005). From 2000 through 2006, the SEC issued 344 Accounting and Auditing Enforcement Releases (AAERs) relating to financial statement fraud, several of which were issued to the same company, and not all investigations resulted in a finding of financial statement fraud.

Corresponding author: Sunita Goel
sgoel@siena.edu
Published Online: December 2010

Due to the nature of financial statement fraud, one of the most difficult tasks in detecting such fraud is the identification of its symptoms. Some of those symptoms might be present even though no fraud exists. Generally Accepted Accounting Principles (GAAP) violations, for example, may not necessarily indicate presence of fraud since departures from GAAP may be appropriate to the company's situation and such departures may have been adequately disclosed. It is possible that only a small number of symptoms may manifest themselves when fraud is occurring because, for example, the fraud may be related to omission by management of disclosures on contingent liabilities or related-party transactions from the notes to the financial statements, and it is difficult to assess their impact before the entire fraud has unraveled. Since symptoms of fraud can be caused by legitimate factors, the mere presence of symptoms cannot necessarily lead to inference of fraud. Moreover, fraud symptoms cannot easily be ranked in order of importance, nor can they be easily combined to yield effective predictive models. Their relative importance varies widely. Fraud detection is muddled by the lack of consensus on symptoms that reliably indicate fraudulent behavior. Nevertheless, it is widely acknowledged that fraud symptoms often exhibit themselves through changes in the financial statements.

The difficulty of detecting fraud is further exacerbated by the fact that financial statements can be misleading even if they are in conformity with GAAP. This is due to the fact that the U.S. GAAP is rules-based and rules cannot be complete in the sense of covering all conceivable situations. It is possible for the companies to be creative in financial measurements as well as disclosures. Therefore, it is necessary to investigate the quantitative information in the financial statements, as well as the qualitative disclosures in the footnotes accompanying the financial statements.

The financial accounting literature is replete with studies that investigate the relationship between the quantitative information in the financial statements and fraud. However, the literature investigating the relationship between the qualitative information accompanying the financial statements and fraud is scant. In this paper, we examine qualitative content of annual reports and explore linguistic features that differentiate fraudulent annual reports from nonfraudulent annual reports.

Most earlier studies of fraud detection have, in our opinion, ignored a key component of financial statements: qualitative textual content in the financial statements. The financial statements communicate quantitative information, qualitative narratives as well as forward-looking information. The disclosures in the qualitative narratives may not contain indications of fraud explicitly, but information regarding fraud, if any, is camouflaged using the rich syntactic as well as semantic arsenal available for writing in natural languages such as English. Indicators of fraud can be constructed from our understanding of such arsenal, and the metrics derived from such indicators can be estimated by the statistical analysis of the qualitative narratives in the financial reports.

In this paper we argue that the textual information released by companies contains indicators in the form of strategically placed phrases, selective use of sentence constructions, selective use of adjectives and adverbial phrases, and similar linguistic variables to conceal fraudulent behavior. An examination of such cues, hidden in the qualitative content of annual reports, can provide new, interesting, and useful information for fraud detection. Systematic and objective statistical analysis of large volumes of text data in the annual reports is important because only a tiny fraction of all corporate information disclosed is quantitative in nature.

The rest of the paper is organized as follows. The next section provides relevant literature on financial statement fraud and fraud detection. This is followed by a description of the sample. We then discuss research methodology and present our results and findings. The last section presents concluding remarks.

RELEVANT LITERATURE

Financial Statement Fraud

Elliot and Willingham (1980) define financial statement fraud as fraud committed by top management through materially misleading financial statements. In Table 1 we summarize the common themes among different definitions in the literature of financial statement fraud (Sawyer 1988; Thornhill and Wells 1993; Arens and Loebbecke 1994; Vanasco 1998; Albrecht et al. 2001) and in the official pronouncements by authoritative bodies (Institute of Internal Auditors [IIA] 1985, 1986; National Commission on Fraudulent Financial Reporting [NCFFR] 1987; Association of Certified Fraud Examiners [ACFE] 1993, 1996).

One can summarize financial statement fraud as an illegitimate act, committed by management, which injures other parties through misleading financial statements. In the literature, the terms "financial statement fraud" and "management fraud" have been used interchangeably (Elliott and Willingham 1980; Robertson 2000) since when financial statement fraud occurs, it typically is with consent or knowledge of management.

Fraud Detection Models

Until recently, most researchers have modeled fraud detection by traditional statistical techniques such as logistic regression (Persons 1995; Beasley 1996; Summers and Sweeney 1998; Lee et al. 1999; Abbot et al. 2000; Bell and Carcello 2000; Spathis 2002), linear discriminant analysis (Fanning and Cogger 1998; Kaminski et al. 2004), and probit analysis (Dopuch et al. 1987; Hansen et al. 1996; Beneish 1999; Lennox 2000). More recently, the studies have used data mining and machine learning techniques to model problems in the domains of accounting and finance. This shift can be attributed to the limitations of the traditional statistical techniques used in the earlier studies. Drawing on the field of Artificial Intelligence (AI), some of the fraud detection models have used neural networks (Green and Choi 1997; Fanning and Cogger 1998), expert systems (Ragothaman et al. 1995; Eining et al. 1997), genetic algorithms (Hoogs et al. 2007), and decision trees (Kirkos et al. 2007) to detect fraud.

A perusal of the above literature shows that most of the studies used financial metrics and ratios extracted from financial statements to detect fraud. Some of these studies have focused on examining the relationship between fraudulent financial reporting and quantitative indicators such as composition of boards of directors, insider trading, auditor rotation, or financial restatements, in addition to financial data.

Furthermore, it should be noted that many studies, including Hansen et al. (1996), Eining et al. (1997), and Bell and Carcello (2000), used internally generated financial information. On the other hand, fraud studies by researchers such as Green and Choi (1997), Summers and Sweeney

TABLE 1

Elements of Financial Statement Fraud

- Intentional conduct, whether by act or omission
- Committed by management
- Results in materially misleading financial statements (which may arise from misrepresentation or omission of material facts)
- Concealment through fraudulent financial reporting (perpetrators have taken steps to hide fraud from others)
- Users of financial statements have relied and acted upon them and in the process have been injured

(1998), Beneish (1999), Kaminski et al. (2004), Hoogs et al. (2007), and Kirkos et al. (2007) showed the benefits of using external information. Summers and Sweeney (1998) demonstrated that their findings hold even when fraud risk factors from prior studies were controlled, indicating an incremental benefit to using external information.

However, the limitations of these models to correctly predict fraud can have serious implications due to high rates of false negatives (Type I error) and false positives (Type II error). Typically, the cost of misclassifying a company involving fraud (i.e., a false negative) is higher than the cost of misclassifying a no-fraud company (i.e., a false positive). For example, if an investor invests in a company that is involved in fraud, but this company has been misclassified as a no-fraud company, he will incur a loss when fraud is discovered. On the other hand, if he does not invest in a no-fraud company as this company is misclassified as a fraud company, he will miss a profitable investment opportunity.

Kaminski et al. (2004) demonstrated the limited ability of financial ratios to detect fraud and concluded that these conventional quantitative financial factors are inadequate for predicting fraud. More recently, Dikmen and Küçükkocaolu (2010) have used a sample of 126 Turkish manufacturing firms described over ten financial ratios to detect factors associated with false financial statements with 82 percent accuracy. Dechow et al. (2011) conducted a detailed analysis of firms investigated by the SEC for misstating quarterly or annual earnings. Using F-ratio, they predicted fraud with 79 percent accuracy.

In contrast, in this paper we use the verbal, qualitative (nonquantitative) content of the annual reports to build our fraud detection model, as textual content of annual reports contains richer information than the financial ratios, which can be easily camouflaged. As our results show, our model performs better than the earlier fraud detection models (see the "Results and Discussion" section).

Qualitative Analysis of Annual Reports

Some prior research highlights the importance of textual portions of annual reports to prime users of financial accounting information such as investors and financial analysts (Abrahamson and Amir 1996; Bryan 1997; Rogers and Grant 1997). For instance, qualitative analysis has been used to predict bankruptcy (Tennyson et al. 1990), financial distress (Boo and Simnett 2002), company performance (Abrahamson and Park 1994), and future viability (Steele 1982). Due to implementation constraints in predicting outcomes such as bankruptcy, financial distress, or company performance, some of these studies utilized only some parts of the annual reports. For example, some of these earlier studies involved manual examination of the qualitative content of annual reports. Manual examination of qualitative content can be very tedious, time-consuming, error-prone, and expensive. Some of the earlier studies also used a hybrid of automated and manual tools to do qualitative analysis, but the researchers limited the use of manual tools to only those portions of annual reports that they suspected to be relevant to their studies. Very few studies have addressed the annual report as a whole, in terms of the integration of the messages across the various parts of the report. In contrast, the research reported here takes advantage of advances in natural language processing, artificial intelligence, and machine learning to examine the entire textual content of annual reports.

None of the previous studies have utilized the qualitative content of annual reports to detect fraud, with the exception of Cecchini (2005). However, our study differs from Cecchini's (2005) work in many significant respects. For example, Cecchini's (2005) qualitative textual analysis for fraud detection was limited to only "Management's Discussion and Analysis of Financial Condition and Results of Operations" (MDNA) section of annual reports and involved examination of the verbal content. In contrast, our study examines verbal content (content words, frequencies of usage, word patterns, etc.) and also the presentation style of the annual reports to explore linguistic

features (such as voice [active versus passive], uncertainty markers, readability index, tone, usage of proper nouns, type-token ratio, etc.) that can distinguish fraudulent annual reports from nonfraudulent reports.

DATA AND SAMPLE SELECTION

The fraud data set in this study consists of companies that were accused of fraudulent financial reporting in the period from 1993 to 2006, i.e., fraud that had affected 10-Ks (annual reports) through material manipulation, misrepresentation, or failure to disclose material facts. Specifically, a company is included in the data set if it was alleged to have violated Rule 10(b)-5 of the Securities Exchange Act of 1934 and, subsequently, sufficient evidence of fraud was found to corroborate such allegations. Rule 10(b)-5 requires the intent to deceive, manipulate, or defraud. Thus, cases where a company was alleged to have accepted kickbacks, to have violated the Foreign Corrupt Practices Act, to have participated in a price-fixing scheme, to have violated antitrust laws, to have conducted wire fraud, to have issued a fraudulent prospectus, or to have committed fraud on registration statements are excluded from our fraud data set.

Fraudulent companies were identified using Lexis-Nexis, Compustat via Research Insight, the *Wall Street Journal* (*WSJ*) Index, and Accounting and Auditing Enforcement Releases (AAERs) issued by the Securities Exchange Commission (SEC) for the period 1993 to 2006. Many empirical studies on financial statement fraud have used the issuance of AAERs as a proxy for financial statement fraud. Even though AAERs provide an objective way of identifying publicly traded companies that have been accused of financial statement fraud, many companies accused in AAERs reach a settlement with the SEC without admitting or denying the allegations, with the result their culpability for fraud is not determinable. In order to make sure that the fraud data set did not include nonfraudulent companies, only those AAERs where companies failed to comply with SEC rules that pertain to fraud and had documented evidence of fraud were considered.

A sample of 126 fraud companies with 405 fraud years identified in the alleged fraud period (1993 to 2006) was selected. Initially, a comprehensive list of all those U.S. publicly listed companies where fraud had occurred and been discovered over the 14-year period from 1993 to 2006 was created. A total of 140 companies were identified during this time period. Out of these 140 companies, 126 companies that had filed their 10-Ks with the SEC and whose 10-Ks were electronically available for download from the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) database were selected. The remaining 14 companies had stopped filing 10-Ks with the SEC due to delisting. Since the 10-Ks of these companies were not available for the fraud period, they were dropped from our fraud data set.

For each fraud company, an attempt was made to select five no-fraud companies. In this study, a peer set for each fraud company, as opposed to a matched-pair data set, was selected, allowing for a data set that approximates a more realistic scenario of the infrequency of fraud. The selection of no-fraud companies was done on the basis of a threefold criterion of year, industry, and size. First, for each fraud company, multiple no-fraud companies were selected for the fraud period. Second, for each fraud company, multiple no-fraud companies were selected from the same industry as defined by two-digit Standard Industrial Classification (SIC) codes, where available, and North American Industry Classification System (NAICS) otherwise. The third criterion for selection of no-fraud companies was that they had to be within the same size range, i.e., within 10–20 percent of the total assets or sales of the fraud company. As a result, 622 U.S. publicly listed corporations with 622 no-fraud years, where fraud had never been reported, were selected for the time period 1993–2006.

In addition, in order to avoid recognizing the company's style, rather than presence of fraud, we did a comparative study on two no-fraud data sets to evaluate the effects of company's style on the performance of the fraud detection model. For this, we created two versions of a no-fraud data

set and each was paired with a fraud data set. As described earlier, the first version of the no-fraud data set consisted of 10-Ks of no-fraud companies that met the selection criteria. In addition to the 10-Ks of no-fraud companies, the second version of the no-fraud data set also included 10-Ks for nonfraudulent years of selected fraud companies that were outside the pre-fraud, fraud, and post-fraud periods. For all the experiments that we conducted with the first version of the no-fraud data set, we repeated each of those experiments with the second version of the no-fraud data set and compared the results.

For the 126 fraud companies, 10-Ks were also collected for pre-fraud years and a post-fraud year ("fraud period–4," "fraud period–3," "fraud period–2," "fraud period–1," "fraud period," "fraud period+1"). The inclusion of four years prior to the fraud period was done to be consistent with the literature findings, which indicate that intensity of fraud grows over time and it usually takes an average of 3.02 years before a fraud is exposed (Summers and Sweeney 1998). Data on the pre-fraud and post-fraud periods were collected to identify features that distinguish early warning signs of fraud from symptoms of advanced fraud in fraudulent companies.

The distribution of data sets for both versions of fraud detection and stages of fraud detection is summarized in Table 2, Panels A and B. The first version of the data set for detecting fraud consisted of 1,027 documents belonging to two categories (fraud, no-fraud). Out of these 1,027 documents, 405 documents were fraudulent 10-Ks of 126 fraud companies, and 622 documents were nonfraudulent 10-Ks of 622 no-fraud companies. The second version of the data set for detecting fraud consisted of 1,375 documents belonging to two categories (fraud, no-fraud). Out of these 1,375 documents, 405 documents were fraudulent 10-Ks of 126 fraud companies, and the remaining 970 documents consisted of 622 nonfraudulent 10-Ks of 622 no-fraud companies and 348 nonfraudulent 10-Ks of 126 fraud companies, which were outside the pre-fraud, fraud, and post-fraud periods.

The first version of the data set for detecting different stages of fraud consisted of 713 documents belonging to three categories (pre-fraud, adv-fraud, post-fraud). Out of these 713 documents, 208 documents were 10-Ks of pre-fraud years of 126 fraud companies, 405 documents were 10-Ks of adv-fraud years of 126 fraud companies, and 100 documents were 10-Ks of post-fraud years of 126 fraud companies. The second version of the data set for detecting different stages of fraud consisted of 613 documents belonging to two categories (pre-fraud, adv-fraud). Out of these 613 documents, 208 documents were 10-Ks of pre-fraud years of 126 fraud companies, and 405 documents were 10-Ks of adv-fraud years of 126 fraud companies.

TABLE 2

Distribution of Data Sets

Panel A: Data Distribution for Fraud Detection

	Version 1	Version 2
Fraud	405	405
No-Fraud	622	970

Panel B: Data Distribution for Detection of Stages of Fraud

	Version 1	Version 2
Pre-Fraud	208	208
Adv-Fraud	405	405
Post-Fraud	100	

For both sets of companies, original 10-Ks were collected and not the restated 10-Ks. The original 10-Ks were selected because a restatement of a financial statement is typically created to correct the previous financial statement for intentional/unintentional errors and accounting irregularities. Restatements represent an acknowledgment by the firm that prior financial statements were not in accordance with Generally Accepted Accounting Principles (Palmrose and Scholz 2004). In order to identify symptoms of fraud and proactively detect fraud, we needed to examine and analyze the original 10-Ks and not the restated 10-Ks.

METHODOLOGY AND RESULTS

The methodology used in this study was implemented using Natural Language Processing (NLP) tools. NLP deals with analyzing, understanding, and generating the language. It includes syntactic, morphological, semantic, and phonological analysis. The application of Natural Language Processing tools for fraud detection is a fertile research area that should be investigated to the fullest possible extent. Unlike research in other well-examined fields of accounting and finance, such as bankruptcy prediction (Zhang et al. 1999; Lensberg et al., 2006), research on detecting financial statement fraud using machine learning-based classifiers such as Naïve Bayes classifiers, neural networks, or support vector machines is a relatively new phenomenon.

In this study, a fraud detection model was built using Support Vector Machines (SVM), a supervised machine learning technique that learns the characteristics (also called features) of positive and negative examples from a training set. Once the learning is successful, SVM is able to successfully classify unlabelled annual reports in the testing data set as fraudulent or nonfraudulent. The correctness of fraud predictions is then evaluated against the correct fraud classes of the testing data set. Several standard evaluation measures such as accuracy, precision, recall, and F-measures are used (Manning and Schütze 1999). These evaluation measures presuppose that each document (annual report) belongs to only a single class (fraudulent or nonfraudulent). The fraud classifier is also trained on pre-fraud and post-fraud data of fraudulent companies to detect early warning signs of fraud.

The methodology presented in this research differs from earlier fraud detection studies using AI techniques as well as non-AI techniques with respect to input vector selection. Most prior studies have selected quantitative information such as financial ratios and metrics as the input vector. Unlike these earlier studies, this study looked at the qualitative factors such as tone, voice, readability index, etc. to assess the likelihood of fraud. In addition, in this research we carried out an in-depth examination of the qualitative content of annual statements in terms of both content and presentation style, unlike some earlier studies where they looked at only one subsection of the annual report to predict bankruptcy, companies' future viability, company performance, and firms' environmental performance.

Results and Discussion: Baseline Approach

For baseline experiments, we used a universally accepted technique for document classification called "bag of words." In a "bag of words" approach, a document is represented with a vector of word counts that appear in it. In this approach, the exact ordering of the words in a document is ignored; instead, information on the number of occurrences of each word is retained. The learning algorithm in this approach examines the "bag of words" vector associated with the incoming document and sees if it fits closely to typical vectors associated with a given class or not. Two documents with similar "bag of words" representations are considered similar in content. Figure 1 illustrates the "bag of words" approach with an example.

For initial baseline experiments, preliminary data preprocessing was conducted in three steps. First, all the words were converted into lower case so that no two same words such as "allege" and "Allege" are included in the corpus as different words. Second, punctuation was removed. Third,

FIGURE 1
Illustration of "Bag of Words" Approach

(a) Snippet from an Annual Report (10-K) of Company A

The Company is a defendant in a number of other pending legal proceedings incidental to present and former operations, acquisitions and dispositions. The Company does not expect the outcome of these proceedings, either individually or in the aggregate, to have a material adverse effect on its financial position, results of operations or liquidity.

(b) Snippet from an Annual Report (10-K) of Company B

In the opinion of the Company, the outcome of any of these matters will not have a material adverse effect on the Company's consolidated financial position but could be material to its results of operations in any one accounting period.

(c) "Bag of Words" for the 10-K snippets shown in (a) & (b)

w1 = the, w2 = company, w3 = is, w4 = a, w5 = defendant, w6 = in, w7 = number, w8 = of, w9 = other, w10 = pending, w11 = legal, w12= proceedings, w13 = incidental, w14 = to, w15 = present, w16 = and, w17 = former, w18 = operations, w19 = acquisitions, w20 = dispositions, w21 = does, w22 = not, w23 = expect, w24 = outcome, w25 = these, w26 = either, w27 = individually, w28 = or, w29 = aggregate, w30 = have, w31 = material, w32 = adverse, w33 = effect, w34 = on, w35 = its, w36 = financial, w37 = position, w38 = results, w39 = liquidity, w40 = opinion, w41 = any, w42 = matters, w43 = will, w44 = companys, w45 = consolidated, w46 = but, w47 = could, w48 = be, w49 = one, w50 = accounting, w51 = period

(d) Document Representation Version of the Partial 10-Ks shown in (a) & (b)

A=(4,2,1,2,1,2,1,3,1,1,1,2,1,2,1,2,1,2,1,1,1,1,1,1,1,1,1,2,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0)

B=(4,1,0,1,0,2,0,4,0,0,0,0,0,1,0,0,0,1,0,0,0,1,0,1,1,0,0,0,0,1,2,1,1,1,1,1,1,1,0,1,2,1,1,1,1,1,1,1,1,1,1)

numbers were removed from all the documents. Since numbers contained in 10-Ks by themselves do not carry discriminative information, the input to the classifiers consisted of only nonnumeric content. Each element of the feature vector space represented a unique word that occurs in the whole corpus, and each nonzero value in a document indicated the number of times a word appeared in that document. This resulted in a very high-dimensional feature space due to a large number of features. At this stage, we did not use a stop words list (removing stop words such as "the," "for," "of," etc.) nor did we perform stemming (converting words to their stems), as we did not want to throw away any useful token. Stemming was not performed as inflected variants of

base forms have very different meanings in the domain of accounting. This is consistent with established practice in existing research (e.g., Chen et al. 1995; Garnsey 2006). We evaluated the stop words separately and adjusted the stop words list so that it did not include any of those words that are relevant for our study. For instance, auxiliary verbs were not included in the stop words list, as these tokens were required to analyze uncertainty marker features. The results of the baseline experiments without applying a stop words list and with an adjusted stop words list are explained in the baseline results section.

Information Gain

Information Gain (IG) is the reduction of entropy with respect to the classification of a target class based on the observation of a feature. In other words, IG indicates how useful a feature is in predicting a class. The basic idea of IG is to retain features that reveal the most information about the distribution of classes. A text feature selection algorithm typically retains words with higher scores and discards words with smaller scores, as words with smaller scores are rarely informative and do not contribute much in prediction of the class. Very often, features whose IG score is less than some predetermined threshold are removed. We use an information gain measure to explore the discriminative power of each unique term and rank the features by the IG score. IG can be computed by subtracting conditional entropy of the class from total entropy of the class. Table 3, Panels A and B, lists the top 25 discriminative words by information gain for detecting fraud and levels of fraud, respectively.

Baseline Results with Naïve Bayes

Naïve Bayes (NB) is one of the simplest and most effective inductive learning algorithms. The basic idea in the NB approach is to use the joint probabilities of words and categories to estimate the probabilities of categories when a document is given (McCallum and Nigam 1998). The NB classifier assigns the most likely class to a given example described by its feature vector. The underlying assumption of the NB approach is that the probability of each word occurring in a document is independent of the occurrence of other words in the document and the probability that a document is generated in some class depends only on the probabilities of the words given

TABLE 3

List of Discriminating Words

Panel A: Ranking of the Top 25 Discriminating Words by Information Gain for Detecting Fraud

1. allege	6. defendants	11. use	16. aggregate	21. colombia
2. argentina	7. manhattan	12. none	17. outstanding	22. seeks
3. brazil	8. cooperating	13. held	18. price	23. shares
4. plaintiffs	9. purported	14. paid	19. taxes	24. plan
5. alleges	10. venezuela	15. about	20. counterparties	25. requirements

Panel B: Ranking of the Top 25 Discriminating Words by Information Gain for Predicting Levels of Fraud

1. sarbanes	6. misleading	11. quantitative	16. llc	21. ethics
2. weaknesses	7. qualitative	12. fraud	17. impaired	22. plaintiff
3. oxley	8. omit	13. certifying	18. com	23. concluded
4. sros	9. eitf	14. fasb	19. defendants	24. dismiss
5. conclusions	10. untrue	15. summarize	20. allege	25. complaint

the context of the class. Even though it is a probabilistic classifier, its classification performance is competitive with the performance of other sophisticated machine learning methods (Mitchell 1997).

For our initial baseline results, we used the Bow (also known as Rainbow) classifier system based on the Bow library, which uses the Naïve Bayes (NB) algorithm as the default algorithm for text classification. Bow is a statistical modeling toolkit for text classification that was developed by Andrew McCallum (1996). Bow has options for both models of Naïve Bayes, i.e., the multivariate Bernoulli model and the multinomial model. For preliminary experiments we use the multinomial model, which has been shown to perform well with large feature sets (McCallum and Nigam 1998).

We applied NB classification to the problem of document categorization, focusing on two issues: (1) fraud detection, and (2) detection of different stages (levels) of fraud. For collecting initial baseline results, we used simple feature reduction techniques (stop words, pruning) with a "bag of words" approach in order to do preliminary exploration of these techniques' potential to improve classification accuracy. Table 4, Panels A and B, shows a comparison of the average classification accuracy rates using ten-fold cross-validation for fraud data sets.

These results indicate that for the first data set, in terms of fraud detection, NB performed best when we applied both pruning and stop words and, for the second data set, NB performed best when we applied pruning only. On the other hand, for the first data set, in terms of detection of

TABLE 4

Baseline Results (NB Classifier)

Panel A: Baseline Results with NB Classifier for Fraud Detection

	Fraud Detection	
Features/Dataset	**Fraud/No-Fraud (Version 1)**	**Fraud/No-Fraud (Version 2)**
Bag of Words (w/o applying stop words and no pruning)	52.17%	52.59%
Bag of Words (with stop words only)	55.28%	53.13%
Bag of Words (with pruning only)	55.11%	53.72%
Bag of Words (with pruning and stop words)	56.75%	51.76%

Panel B: Baseline Results with NB Classifier for Detection of Stages of Fraud

	Detection of Stages of Fraud	
Features/Dataset	**Pre/Adv/Post (Version 1)**	**Pre/Adv (Version 2)**
Bag of Words (w/o applying stop words and no pruning)	42.54%	51.14%
Bag of Words (with stop words only)	40.28%	50.97%
Bag of Words (with pruning only)	41.97%	54.58%
Bag of Words (with pruning and stop words)	39.01%	51.79%

stages of fraud, NB performed best when we neither applied pruning nor stop words and its performance improved as the number of features increased. For the second data set of detection of stages of fraud, NB performed best when we applied pruning only.

For detecting different stages of fraud, our initial baseline results showed that when the classifier was trained and tested on data of three classes (pre-fraud, adv-fraud, and post-fraud), its best performance score was 42.54 percent. On the other hand, when we trained and tested the classifier with data of only two classes (pre-fraud and adv-fraud), its performance increased to 54.58 percent. The analysis of classifier errors in the set of three classes indicated that the classifier misclassified all instances in the post-fraud class by assigning 90 percent of the instances to the adv-fraud class and 10 percent of the instances to the pre-fraud class. Most of these misclassifications seemed to occur due to the large overlap of terms found between the annual reports issued during the fraud period and the post-fraud period. Another likely reason for poor performance of the classifier was that the size of the training data set in the minority categories such as post-fraud was too small to provide adequate training data. This was our motivation to collapse categories in the second version of the data set relating to detection of stages of fraud.

The reasons for using the Naïve Bayes classifier for our initial baseline results were: (1) it is easy to implement, (2) it is among the most successful known algorithms after SVM for text classification (Dumais et al. 1998), (3) the experimental results helped us compare the performance of the two popular text classifiers (NB and SVM) and back our claims that SVM is better suited to our problem of fraud detection, and (4) it acted as a preprocessor to explore useful feature subsets for SVM.

Baseline Results with Support Vector Machines

In the previous subsection, we presented preliminary baseline results using the Naïve Bayes classifier. We also ran baseline experiments using Support Vector Machines (SVM), the main classifier used in this study. SVM is a supervised machine learning technique that is based on statistical learning theory. The SVM algorithm learns by example to classify objects into a fixed number of predefined categories. In this study, the SVM was trained to recognize fraudulent annual reports by examining hundreds of fraudulent and nonfraudulent annual reports. SVMs are based on the Structured Risk Minimization (SRM) principle drawn from computational learning theory (Vapnik and Cheronekis 1974). SRM is an inductive principle for model selection that provides a trade-off between hypothesis space complexity and the quality of fitting the training data that guarantees the lowest true error on an unseen and randomly selected test example. SVMs determine a hyperplane in the feature space that best separates positive from negative examples.

We used Waikato Environment for Knowledge Analysis (WEKA) in our experiments to train the SVM classifier to build the fraud detection model. WEKA is a machine learning toolkit that supports data mining tasks such as classification, clustering, and regression, and contains visualization tools for data analysis, data preprocessing, and feature selection (Witten and Frank 2005).

It should be noted that unlike Bow (toolkit used to train Naïve Bayes classifier), WEKA only takes data files that are in Attribute Relation File Format (ARFF) format. Thus, 10-Ks that were downloaded from EDGAR and saved as text files could not be submitted to WEKA for processing in the raw format. Therefore, we converted these files into ARFF format before feeding them to WEKA for processing. Figure 2 shows a sample of an ARFF data file where a stop words list has not been applied. This file contains 261,110 features (words) and 1027 instances (405 fraudulent and 622 nonfraudulent documents). Due to the large number of features, it is not possible to show the entire ARFF file.

Our baseline results with the SVM classifier for detecting fraud and detecting different stages of fraud are presented in Table 5, Panels A and B. As discussed earlier, in this case also, we used ten-fold cross-validation to train the SVM classifier. These results indicate that SVM performed

FIGURE 2
Sample of ARFF File Used in WEKA

```
@relation fraud

@attribute communications numeric
@attribute corporation numeric
@attribute cable numeric
@attribute of numeric
@attribute contents numeric
@attribute part numeric
@attribute item numeric
@attribute business numeric
@attribute properties numeric
@attribute legal numeric
@attribute proceedings numeric
@attribute submission numeric
.
.
.
.
.
.
@attribute Indicator {Fraud, NoFraud}
@data
295,281,24,12641,12,157,66,880,76,88,88,4,……..…,Fraud
9,410,9,4743,8,55,36,337,21,44,18,3,……...,Fraud
3,54,37,2457,0,22,29,101,4,9,2,1,……..Fraud
5,290,21,9311,6,134,65,388,46,64,36,6,……...,Fraud
2,268,23,5528,7,84,62,149,173,69,25,6,……Fraud
.
.
.
.
.
22,13,15,1449,1,31,63,36,4,11,13,2,…….,NoFraud
18,43,5,831,0,18,20,31,12,11,6,…….,NoFraud
20,7,5,773,1,22,31,36,2,3,3,2,……...…,NoFraud
1,46,80,5650,68,133,68,232,82,65,44,6,……..…..,NoFraud
22,58,1,1235,0,19,25,133,2,9,6,2,……..…,NoFraud
.
.
```

best in most of the data sets when we applied both pruning and the adjusted stop words list in our experiments. It can be observed that our baseline results based on a "bag of words" approach were much better with SVM than those achieved with the Naïve Bayes classifier.

When baseline results for fraud detection were compared to the random baseline, we noted that baseline experiments with SVM were able to beat the random baseline, whereas Naïve Bayes

TABLE 5

Baseline Results (SVM Classifier)

Panel A: Baseline Results with SVM Classifier for Fraud Detection

	Fraud Detection	
Features/Dataset	**Fraud/No-Fraud (Version 1)**	**Fraud/No-Fraud (Version 2)**
Bag of Words (w/o applying stop words and no pruning)	63.11%	63.01%
Bag of Words (with stop words only)	66.09%	67.78%
Bag of Words (with pruning only)	66.76%	66.91%
Bag of Words (with pruning and stop words)	71.67%	69.11%

Panel B: Baseline Results with SVM Classifier for Detection of Stages of Fraud

	Detection of Stages of Fraud	
Features/Dataset	**Pre/Adv/Post (Version 1)**	**Pre/Adv (Version 2)**
Bag of Words (w/o applying stop words and no pruning)	51.81%	62.79%
Bag of Words (with stop words only)	52.41%	62.87%
Bag of Words (with pruning only)	52.37%	65.73%
Bag of Words (with pruning and stop words)	51.83%	65.81%

Panel C: Detailed Results for Fraud Detection Dataset with the Best Accuracy of 71.67%

Class	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-measure**
Fraud	0.415	0.087	0.757	0.415	0.536
No-Fraud	0.913	0.585	0.706	0.913	0.796

was unable to beat the random baseline. For fraud detection, the random baseline would yield an accuracy of 60.56 percent simply by classifying all the documents to the largest class (no-fraud).

In addition, we report detailed accuracy results by class in Table 5, Panel C. It shows the True Positive (TP) rate, False Positive (FP) rate, precision, recall, and F-measure results for the data set with the best score of 71.67 percent. For the fraud class, true positives indicate the number of fraudulent 10-Ks that are correctly classified as fraudulent, whereas false positives indicate the number of nonfraudulent 10-Ks that are incorrectly classified as fraudulent. For the no-fraud class, true positives indicate the number of nonfraudulent 10-Ks that are correctly classified as nonfraudulent, and false positives indicate the number of fraudulent 10-Ks that are incorrectly classified as nonfraudulent. The TP rate is obtained by dividing the number of true positives by the sum of true positives and false negatives. The FP rate is obtained by dividing the number of false positives by the sum of true negatives and false positives. Precision for a class is defined as the number of correct predictions of a class divided by the total number of predictions for that class, whereas recall for a class is defined as the number of correct predictions of a class divided by the total number of actual instances of that class in the data set. The F-measure is the weighted harmonic mean of precision and recall; it is a measure of the performance of the classifier.

These results indicate that the FP rate of 0.585 for the no-fraud class is much higher than the FP rate of 0.087 for the fraud class. The high rate of false positives for the no-fraud class is not a desirable situation, as this indicates that the classifier missed 58.5 percent of the fraudulent annual

reports (Type I error) and misclassified them as nonfraudulent annual reports, which is more dangerous than the case where the classifier misclassifies nonfraudulent annual reports as fraudulent annual reports (Type II error), thus creating a false alarm.

These results also indicate that the TP rate of 0.415 for the fraud class is much lower than the TP rate of 0.913 of the no-fraud class. Here, the low TP rate of 0.415 for the fraud class is not a desirable situation, even though its FP rate is low. In general, the FP rate goes up as one attempts to increase the TP rate. This is evident in the case of the no-fraud class, which has a high TP rate of 0.913 along with high FP rate of 0.585. Here, the classifier was more liberal in the sense that it made positive predictions even when there was weak evidence, which resulted in a high TP rate but also a high FP rate.

As observed in Table 5, Panel C, even though the recall rate for the minority class "Fraud" was lower (which might be due to the fact that the distribution of data is skewed), its predictive accuracy was higher. This is due to the fact that the classifier got only those instances of fraudulent annual reports correct where strong evidence was present; therefore, there is a low TP rate for the fraud class, but also few false positive errors. From our baseline experiments, we conclude that further training of the classifier is required with more sophisticated features to minimize the FP rate for both no-fraud and fraud classes, as well as to maximize the TP rate for the fraud class. We discuss the next series of these experiments in subsection "Style and Content Features."

Style and Content Features

Our initial baseline results with a Naïve Bayes classifier, using a "bag of words" approach, were modest, correctly classifying about 56.75 percent. However, when we used Support Vector Machines (SVM) as the main classifier, the fraud classification accuracy, even with baseline features, increased to 71.67 percent.

Inspired by encouraging baseline results, we ran our classifier experiments with four feature sets to examine both the verbal content and the presentation style of the annual reports to detect fraud and stages of fraud. The features relating to content focus on the "what" part of the annual report, that is, what it contains, whereas features relating to presentation style focus on the "how" part of the annual report, that is, how its content is communicated.

We primarily used two tools—DICTION 5.0 and STYLE—for extracting most of our linguistic features. DICTION 5.0 is a Windows-based commercially available text analysis program created by Roderick Hart (2000). STYLE is a UNIX command-line-based GNU program that analyzes the surface characteristics of the writing style of a document, including sentence length and type, word usage, and other readability measures, and provides a stylistic profile of writing at the word and sentence level (Cherry and Vesterman 1991).

The first feature set consists of eight simple surface features such as the average length of the words, the standard deviation of the word length, the average length of the sentences, the standard deviation of the sentence length, the percentage of short and long sentences, the average length of the paragraphs, and the standard deviation of the paragraph length in the document (see Table 6). Even though these features are often called simple surface features, their importance is undeniable. Prior research in the area of stylometry (Forsyth and Holmes 1996) has reported good results with several of these features.

The second feature set consists of four features: voice, frequency of uncertainty markers, tone, and readability. Voice is that form of a verb which shows whether what is denoted by the subject of the sentence does something (active) or has something done to it (passive). Uncertainty markers (also known as hedge words or modal verbs) include words such as "shall," "may," "probably," "possibly," "might," etc. Uncertainty markers have been extensively used in the literature to study style, expression, affect, and attitude in text (Lackoff 1973; Glover and Hirst 1996; Uzuner and Katz 2005a; Rubin et al. 2006). Several studies relating to deception analysis have also used

TABLE 6

Simple Surface Features for a Sample Fraudulent and Nonfraudulent Annual Report

Surface Features	Fraudulent 10-K	Nonfraudulent 10-K
Average Length of the Words (in terms of characters)	5.36	4.93
Standard Deviation of Word Lengths	1.7	1.5
Average Length of the Sentences (in terms of words)	31.1	17.5
Standard Deviation of Sentence Lengths	11.2	5.3
Percentage of Short Sentences (at most 30 words)	60%	55%
Percentage of Long Sentences (at least 60 words)	22%	14%
Average Length of the Paragraphs (in terms of sentences)	3.7	2.3
Standard Deviation of Paragraph Lengths	1.9	1.4

uncertainty markers to isolate cases of deception. The tone defines the semantic orientation of a text and can be measured by examining the lexical choices made by the writer, i.e., words chosen to indicate polarity of the tone. For this, we develop two categories of tone—positive and negative—based on the prior work of researchers in this context (Abrahamson and Amir 1996; Smith and Taffler 2000; Henry 2006). The original list of positive and negative words was adjusted with words found in the fraud corpus. Another feature that we examined in the second feature set was readability. Readability indices are measures of the ease or difficulty of reading and understanding a piece of text. There are different measures available to compute readability grades, such as "Flesch-Kincaid Grade Level," "Automated Readability Index," "Coleman-Liau Index," "Flesch Reading Ease Score," "Gunning Fog Index," "Lix Formula," and "SMOG-Grading." Several studies (Courtis 1986; Baker and Kare 1992; Smith and Taffler 1992; Subramanian et al. 1993) have used readability tests to examine the relationship between the readability of annual reports and corporate failures or corporate profitability. Many studies in computational linguistics have also used readability indices to examine readability of texts (see, for example, Mikk 1995; Das and Roychoudhury 2006). We compared the scores on these readability grades for fraudulent 10-Ks and nonfraudulent 10-Ks to detect fraud.

The third feature set involves deeper linguistic analysis and consists of 14 features. Prior research suggests that markers of linguistic style—articles, pronouns, prepositions, and conjunctions—are, in many respects, as meaningful as specific nouns and verbs in telling what people are thinking and feeling (Dulaney 1982; Colwell et al. 2002; Zhou et al. 2002). In deeper surface features, we examined the vocabulary frequencies (of proper nouns, pronouns, conjunctions, prepositions, nominalizations, verb types, sentence openers) in addition to the vocabulary richness (type-token ratio) in order to explore the underlying grammatical relations and identify patterns of usage in writings of the two corpora (fraud, no-fraud). Zhou et al. (2002) found that high variety index (type-token ratio) is associated with deception. They noted that in cases of deception, the writer uses superfluous and meaningless language to give the impression of completeness. Simple surface features examined earlier indicated that there were structural differences between fraudulent and nonfraudulent annual reports. However, the features examined in the simple surface feature set are under the conscious control of the writers. On the contrary, Yule (1938) found that some of the useful features that represent the specific style are those that the writer does unconsciously. Holmes (1994) noted that features such as the use and frequency of function words (determiners, conjunctions, and prepositions, etc.) were useful for characterizing

the style, as they were not under the conscious control of the writer. We believed that investigation of these features would help us in distinguishing simple styles of annual reports from the ponderous ones.

The fourth feature set consists of content-related features such as keywords, bigrams, and TFIDF words. In our keyword-based approach to classification, we selected the top 100 words with the highest information gain in the training corpus as fraud indicators. Bigrams are a special case of n-grams and can consist of a sequence of two characters, two words, or two syllables. Thus, in the case of bigrams, the feature vector consists of pairs of words instead of single words, such as "sarbanes oxley" and "generally accepted." We extracted bigrams collocating with the keywords. Finally, we used the Weighted-Term Frequency Inverse Document Frequency (TFIDF) measure to evaluate how central a content word is in the fraud and no-fraud corpus. There is extensive literature demonstrating that frequencies of words convey information regarding their importance and content captures otherwise hard-to-quantify concepts (see, for example, Zipf 1929, 1949; Luhn 1957; Iker 1974; Weber 1990; Gangolly and Wu 2000; Hand et al. 2001; Uzuner and Katz 2005b).

The pre-selection of these features was inspired by our informed reasoning and domain knowledge, and rests on the speculation that the qualitative content of annual report manifests linguistic cues that can be used for detecting fraud. As the classifier converged to higher levels of accuracy, we isolated features that had the most discriminative power in terms of detecting fraud and stages of fraud and ranked them in order of their relevance to our domain problem.

Feature Selection

Feature selection is a common technique that is used in machine learning to select a subset of relevant features from available potential candidate features. In this study, we used a forward feature stepwise selection approach to feature selection and incrementally added the features, one at a time, in the feature space. This way, we were able to understand the effect of different features on the classifier performance and construct a feature set that was most relevant for fraud detection and detection of different stages of fraud.

Some of the core methods used for feature selection are document frequency, information gain, mutual information, and Chi-square. Document frequency counts the number of documents containing the feature. Information gain is the number of bits of information obtained for category prediction given a feature. Mutual information measures mutual dependence of the two variables. Chi-square measures the lack of independence between a term and the category.

In the case of fraud detection, we used the Chi-square method to select features that show statistically significant differences between the fraudulent and nonfraudulent annual reports. For detection of stages of fraud, we also used Chi-square to select features that show statistically significant differences between the annual reports of pre-fraud and adv-fraud periods. Chi-square feature selection has been shown to not only reduce the feature space effectively by reducing the noise introduced in the classifier, but also to improve performance of the classifier at the same time.

Top Ranking Features

Our final set of selected features included only those features that played a role in increasing the overall accuracy of the fraud classifier. Tables 7 and 8 provide a ranking of the top ten features for recognizing fraudulent and nonfraudulent annual reports and for recognizing different stages of fraud, respectively. The features that did not contribute to classifier performance (with the highest p-value), such that their inclusion made no difference in the classifier accuracy, were also eliminated from the feature space.

TABLE 7

Top Ten Features for Detecting Fraudulent and Nonfraudulent Annual Reports

Rank	Feature
1	Percentage of Passive-Voice Sentences
2	Percentage of Active-Voice Sentences
3	Standard Deviation of Sentence Lengths
4	Readability Index
5	Scaled Frequency of Uncertainty Markers
6	Percentage of Sentences Beginning with Subordinating Conjunction
7	Type-Token Ratio
8	Scaled Frequency of Proper Nouns
9	Percentage of "To Be" Verbs
10	TFIDF Weighted Tokens

TABLE 8

Top Ten Features for Detecting Stages of Fraud ("Pre-Fraud," "Advanced-Fraud")

Rank	Feature
1	Readability Index
2	Percentage of Passive-Voice Sentences
3	Percentage of Active-Voice Sentences
4	Standard Deviation of Sentence Lengths
5	Scaled Frequency of Uncertainty Markers
6	Type-Token Ratio
7	Percentage of Words that Belong to "Positive" Tone Category
8	Standard Deviation of Word Lengths
9	Percentage of Sentences Beginning with Subordinating Conjunction
10	Scaled Frequency of Proper Nouns

Fraud Model Results

Our fraud classifier results with the highest ranked features yielded an accuracy of 89.51 percent, which was much higher than our baseline results of 71.67 percent obtained using the "bag of words" approach. Table 9, Panel A, presents average scores of classifier accuracy over ten-fold cross-validation. Here, we also report the detailed accuracy results for Version 1 of the fraud detection data set and for detection of stages of fraud data set in Table 9, Panels B and C, respectively.

These results support our claim that annual reports contain linguistic cues that can be exploited to proactively detect fraud. Furthermore, these results suggest that the subset of features we have selected for fraud detection can be used successfully to distinguish fraudulent annual reports from nonfraudulent annual reports 89.51 percent of the time. Similarly, the subset of features that we have selected for detection of stages of fraud can be used successfully to distinguish early symptoms of fraud from advanced stages of fraud 87.98 percent of the time.

For the best score of 89.51 percent, these results indicate that the TP rate of 0.899 for the fraud class is higher than the TP rate of 0.894 for the no-fraud class. These results also indicate that the FP rate of 0.101 for the no-fraud class is lower than the FP rate of 0.106 for the fraud class. This means that the classifier missed only 10.1 percent of the fraudulent annual reports

TABLE 9

Final Results

Panel A: Average Classifier Accuracy for the Three Data Sets with Most Useful Features

Datasets	Average Classifier Accuracy
Fraud Detection Version 1	89.51%
Fraud Detection Version 2	89.04%
Detection of Stages of Fraud	87.98%

Panel B: Detailed Results for Fraud Detection Version 1 Data Set with the Best Accuracy of 89.51%

Class	TP Rate	FP Rate	Precision	Recall	F-measure
Fraud	0.899	0.106	0.847	0.899	0.872
No-Fraud	0.894	0.101	0.931	0.894	0.912

Panel C: Detailed Results for Detection of Stages of Fraud Data Set with the Best Accuracy of 87.98%

Class	TP Rate	FP Rate	Precision	Recall	F-measure
Adv	0.884	0.130	0.930	0.884	0.906
Pre	0.870	0.116	0.794	0.870	0.830

(Type I error) and misclassified them as nonfraudulent annual reports, whereas the classifier misclassified nonfraudulent annual reports as fraudulent annual reports (Type II error) 10.6 percent of the time.

When comparing performance of the classifier with top ten features to its performance with baseline features, we notice that for the no-fraud class, the TP rate of 0.894 is lower than the TP rate of 0.913, which was obtained with baseline features. However, it should be noted that this is a more desirable situation for the no-fraud class, as its FP rate has tremendously gone down from 0.585, obtained with baseline features, to 0.101. As observed in Table 9, Panel B, the recall rate for the minority class fraud has gone up to 0.899 from 0.415, achieved with the baseline experiments, and its predictive accuracy has also increased to 0.847 from 0.757. This indicates that the classifier was able to overcome the class imbalance problem as it converged to higher levels of learning. Thus, our intuition was correct that as the classifier was trained with more sophisticated features, the learning accuracy of the classifier increased even for the minority class fraud. In addition, our classifier results for detection of stages of fraud with the highest ranked features yielded an accuracy of 87.98 percent, which is much higher than our baseline results of 65.81 percent. When these results for fraud detection were compared to the random baseline, we noted that the classifier was able to beat the random baseline by a much wider margin.

CONCLUSION

In this research, we presented a methodology that involved linguistic analysis of the textual content of annual reports for detecting fraud. Linguistic cues not only helped us in interacting with the text, but also allowed us to look beyond the content of the annual reports. By doing both stylistic analysis and content analysis of these annual reports, we were able to build a fraud detection model that is competitive with the leading fraud detection models and achieves very good results in terms of precision and recall.

The results of our study suggest that the qualitative narrative content of annual reports contains information that is useful for detecting fraud that is not accurately captured by financial

numbers. We found systematic differences in communication and writing style of fraudulent annual reports. For example, fraudulent annual reports contained more passive-voice sentences, used more uncertainty markers, had a higher type-token ratio (lexical variety), and were more difficult to read and comprehend than nonfraudulent annual reports.

To our knowledge, no prior studies have used linguistic features to examine the qualitative content of annual reports to detect fraud. The linguistic differences reported in fraudulent and nonfraudulent annual reports are not meant to oversimplify issues of detecting fraud, but to provide insight and understanding into the ways the companies portray themselves that require further investigation. Our work in the area of fraud detection using linguistic analysis has opened up many possibilities by adding another dimension to corporate fraud research. Our fraud model can be used for predicting early warning signs of forthcoming accounting problems in potentially fraudulent companies and can be of interest to practitioners such as auditors, fraud examiners, and analysts.

This study is subject to several limitations. Our fraud data set consists of fraud companies that have documented evidence of fraud. These companies might not be representative of all companies that have committed fraud. This is due to the fact that companies that commit fraud are less forthcoming about this information (Higson 1999). Like most other fraud studies, companies with undetected frauds are not included in the fraud sample. For the same reason mentioned above, no-fraud companies may also include companies where fraud had occurred but it has not been publicly discovered. Another limitation relates to supervised learning algorithms, that they cannot discover a novel feature unless it is either learned from the training data set or defined by a user. In addition, if a data set is imbalanced, then SVMs tend to produce a less effective classification boundary skewed to the minority class. When there are too few positive examples, SVMs may totally fail, as there is insufficient evidence for statistical learning. Finally, findings of our research are limited, as we examine only annual reports. Future research could examine quarterly reports, which may provide additional insights for fraud detection. Another direction for future research in the area of fraud detection can be to create a fraud ontology, which not only covers corporate fraud but also other types of frauds.

REFERENCES

Abbot, J. L., Y. Park, and S. Parker. 2000. The effects of audit committee activity and independence on corporate fraud. *Managerial Finance* 26 (11): 55–67.

Abrahamson, E., and E. Amir. 1996. The information content of the president's letter to shareholders. *Journal of Business Finance & Accounting* 23 (8): 1157–1182.

———, and C. Park. 1994. Concealment of negative organizational outcomes: An agency theory perspective. *Academy of Management Journal* 37 (5): 1302–1334.

Albrecht, C. C., W. S. Albrecht, and J. G. Dunn. 2001. Conducting a pro-active fraud audit: A case study. *Journal of Forensic Accounting* II: 203–218.

Arens, A., and J. Loebbecke. 1994. *Auditing: An Integrated Approach*. 6th edition. Englewood Cliffs, NJ: Prentice-Hall.

Association of Certified Fraud Examiners (ACFE). 1993. *Cooking the Books: What Every Accountant Should Know About Fraud*. No. 92-5401. Austin, TX: ACFE.

———. 1996. *Report to the Nation on Occupational Fraud and Abuse*. The Wells Report. Austin, TX: ACFE.

Baker, H. E., III, and D. D. Kare. 1992. Relationship between annual report readability and corporate financial performance. *Management Research News* 15: 1–4.

Beasley, M. S. 1996. An empirical analysis of the relation between the board of director composition and financial statement fraud. *The Accounting Review* 71 (4): 443–465.

Bell, T. B., and J. V. Carcello. 2000. A decision aid for assessing the likelihood of fraudulent financial reporting. *Auditing: A Journal of Practice & Theory* 19 (1): 169–175.

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

Boo, E., and R. Simnett. 2002. The information content of management's prospective comments in financially distressed companies: A note. *Abacus* 38 (2): 280–295.

Bryan, S. H. 1997. Incremental information content of required disclosures contained in management discussion and analysis. *The Accounting Review* 72 (2): 285–301.

BusinessWeek. 2005. AIG: What went wrong. A look at how the icon of insurance got itself in such a mess—And where all probes are headed. Available at: http://www.businessweek.com/magazine/content/05_15/b3928042_mz011.htm.

Cecchini, M. 2005. Quantifying the risk of financial events using kernel methods and information retrieval. Doctoral dissertation, University of Florida.

Chen, H., T. Yim, and D. Fye. 1995. Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science* 46 (3): 175–193.

Cherry, L. L., and W. Vesterman. 1991. *Writing Tools—The STYLE and DICTION Programs. In 4.3BSD UNIX System Documentation*. Berkley, CA: University of California.

Colwell, K., C. Hiscock, and A. Memon. 2002. Interviewing techniques and the assessment of statement credibility. *Applied Cognitive Psychology* 16: 287–300.

Courtis, J. K. 1986. An investigation into annual report readability and corporate risk return relationships. *Accounting and Business Research* (Autumn): 285–294.

Das, S., and R. Roychoudhury. 2006. Readability modelling and comparison of one and two parametric fit: A case study in Bangla. *Journal of Quantitative Linguistics* 13 (1): 17–34.

Dikmen, B., and G. Küçükkocaolu. 2009. The detection of earnings manipulation: The three-phase cutting plane algorithm using mathematical programming. *Journal of Forecasting* 29 (5): 442–466.

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

Dopuch, N., R. Holthausen, and R. Leftwich. 1987. Predicting audit qualifications with financial and market variables. *The Accounting Review* 62 (3): 431–454.

Dulaney, E. 1982. Changes in language behavior as a function. *Human Communication Research* 9 (1): 75–82.

Dumais, S., J. Platt, D. Heckerman, and M. Sahami. 1998. Inductive learning algorithms and representations for text categorization. In *Proceedings of the 7th International Conference on Information and Knowledge Management*, 148–155. New York, NY: ACM Press.

Eining, M. M., D. R. Jones, and J. K. Loebbecke. 1997. Reliance on decision aids: An examination of auditors' assessment of management fraud. *Auditing: A Journal of Practice & Theory* 16 (2): 1–19.

Elliott, R. K., and J. J. Willingham. 1980. *Management Fraud: Detection and Deterrence*. New York, NY: Petrocelli Books.

Fanning, K., and K. Cogger. 1998. Neural network detection of management fraud using published financial data. *International Journal of Intelligent Systems in Accounting, Finance and Management* 7: 21–24.

Forsyth, R. S., and D. I. Holmes. 1996. Feature-finding for text classification. *Literary and Linguistic Computing* 11 (4): 163–174.

Gangolly, J., and Y. Wu. 2000. On the automatic classification of accounting concepts: Preliminary results of the statistical analysis of term-document frequencies. *The New Review of Applied Expert Systems and Emerging Technologies* 6: 81–88.

Garnsey, M. R. 2006. Automatic classification of financial accounting concepts. *Journal of Emerging Technologies in Accounting* 3: 21–39.

Glover, A., and G. Hirst. 1996. Detecting stylistic inconsistencies in collaborative writing. In *The New Writing Environment: Writers at Work in a World of Technology*, edited by Sharples, M., and T. Geest. London, U.K.: Springer-Verlag Company.

Green, B. P., and J. H. Choi. 1997. Assessing the risk of management fraud through neural-network technology. *Auditing: A Journal of Practice & Theory* 16: 14–28.

Hand, D., H. Mannila, and P. Smyth. 2001. *Principles of Data Mining*. Cambridge, MA: The MIT Press.

Hansen, J. V., J. B. McDonald, W. F. Messier, and T. B. Bell. 1996. A generalized qualitative-response model and the analysis of management fraud. *Management Science* 42 (7): 1022–1033.

Hart, R. P. 2000. *Diction 5.0: The Text Analysis Program*. Computer Software. Thousand Oaks, CA: Sage.

Henry, E. 2006. Market reaction to verbal components of earning press releases: Event study using a predictive algorithm. *Journal of Emerging Technologies in Accounting* 3: 1–19.

Higson, A. 1999. Why is management reticent to report fraud? An exploratory study. *22nd Annual Congress of European Accounting Association*. Bordeaux, France: European Accounting Association.

Holmes, D. 1994. Authorship attribution. *Computers and the Humanities* 28: 87–106.

Hoogs, B., T. Kiehl, C. Lacomb, and D. Senturk. 2007. A genetic algorithm approach to detecting temporal patterns indicative of financial statement fraud. *International Journal of Intelligent Systems in Accounting, Finance and Management* 15: 41–56.

Iker, H. 1974. An historical note on the use of word-frequency contiguities in content analysis. *Computers and the Humanities* 8: 93–98.

Institute of Internal Auditors (IIA). 1985. *Deterrence, Detection, Investigation, and Reporting of Fraud*. Altamonte Springs, FL: IIA.

———. 1986. *The Role of Internal Auditors in the Deterrence, Detection and Reporting of Fraudulent Financial Reporting*. Altamonte Springs, FL: IIA.

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

Kirkos, E., C. Spathis, and Y. Manolopoulos. 2007. Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications* 32: 995–1003.

Lackoff, G. 1973. Hedges: A study of meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic* 2 (4): 458–508.

Lee, T., R. Ingram, and T. Howard. 1999. The difference between earnings and operating cash flow as an indicator of financial reporting fraud. *Contemporary Accounting Research* 16: 749–786.

Lennox, C. 2000. Do companies successfully engage in opinion-shopping? Evidence from the U.K. *Journal of Accounting and Economics* 29 (3): 321–337.

Lensberg, T., A. Eilifsen, and T. E. McKee. 2006. Bankruptcy theory development and classification via genetic programming. *European Journal of Operational Research* 169: 677–697.

Luhn, H. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development* 1 (4): 309–317.

Manning, C. D., and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

McCallum, A. 1996. *Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering*. Available at: http://www.cs.cmu.edu/~mccallum/bow.

———, and K. Nigam. 1998. A comparison of event models for naive Bayes text classification. In *Proceedings of AAAI/ICML-98 Workshop on Learning for Text Categorization*, 41–48. Madison, WI: AAAI Press.

Mikk, J. 1995. Methods of determining optimal readability of texts. *Journal of Quantitative Linguistics* 2: 125–132.

Mitchell, T. M. 1997. *Machine Learning*. New York, NY: McGraw-Hill.

National Commission on Fraudulent Financial Reporting (NCFFR). 1987. *Report of the National Commission on Fraudulent Financial Reporting*. New York, NY: NCFFR. Available at: http://www.coso.org/Publications/NCFFR.pdf.

Palmrose, Z-V., and S. Scholz. 2004. The accounting causes and legal consequences of non-GAAP reporting: Evidence from restatements. *Contemporary Accounting Research*: 21 (1): 139–180.

Persons, O. 1995. Using financial statement data to identify factors associated with fraudulent financial reporting. *Journal of Applied Business Research* 11: 38–46.

Ragothaman, S., J. Carpenter, and T. Buttars. 1995. Using rule induction for knowledge acquisition: An expert systems approach to evaluating material errors and irregularities. *Expert Systems with Applications* 9 (4): 483–490.

Robertson, J. C. 2000. *Fraud Examination for Managers and Auditors*. Austin, TX: Viesca Books.

Rogers, R. K., and J. Grant. 1997. Content analysis of information cited in reports of sell-side financial analysts. *Journal of Financial Statement Analysis* 3 (1): 14–30.

Rubin, V. L., E. D. Liddy, and N. Kando. 2006. Certainty identification in texts: Categorization model and manual tagging results. In *Computing Attitude and Affect in Text: Theory and Applications (The*

Information Retrieval Series), edited by Shanahan, J. G., Y. Qu, and J. Wiebe, 61–76. New York, NY: Springer-Verlag Company.

Sawyer, L. 1988. *Internal Auditing*. Altamonte Springs, FL: The Institute of Internal Auditors.

Smith, M., and R. Taffler. 1992. The chairman's statement and corporate financial performance. *Accounting and Finance* 32 (2): 75–90.

———, and ———. 2000. The chairman's statement: A content analysis of discretionary narrative disclosures. *Accounting, Auditing & Accountability Journal* 13 (5): 624–646.

Spathis, C. 2002. Detecting false financial statements using published data: Some evidence from Greece. *Managerial Auditing Journal* 17: 179–191.

Steele, A. 1982. The accuracy of chairman's non-quantified forecasts: An exploratory study. *Accounting and Business Research* (Summer): 215–230.

Subramanian, R., R. G. Insley, and R. D. Blackwell. 1993. Performance and readability: A comparison of annual reports of profitable and unprofitable corporations. *Journal of Business Communication* 30: 49–61.

Summers, S. L., and J. T. Sweeney. 1998. Fraudulently misstated financial statements and insider trading: An empirical analysis. *The Accounting Review* 73 (1): 131–146.

Tennyson, B. M., R. W. Ingram, and M. T. Dugan. 1990. Assessing the information content of narrative disclosures in explaining bankruptcy. *Journal of Business Finance & Accounting* 17 (3): 390–410.

Thornhill, W. T., and J. T. Wells. 1993. *Fraud Terminology Reference Guide*. Austin, TX: Association of Certified Fraud Examiners.

Uzuner, O., and B. Katz. 2005a. Capturing expression using linguistic information. In *Proceedings of the 20th National Conference on Artificial Intelligence*. Available at: http://people.csail.mit.edu/ozlem/aaai05UzunerO.pdf.

———, and ———. 2005b. Style versus expression in literary narratives. In *Proceedings of the 28th Annual International ACM SIGIR Conference*. Available at: http://people.csail.mit.edu/ozlem/sigir-05-cc-UzunerO-cr.pdf.

Vanasco, R. R. 1998. Fraud auditing. *Managerial Auditing Journal* 13: 4–71.

Vapnik, V., and A. Chervonenkis. 1974. *Theory of Pattern Recognition*. Moscow, Russia: Nauka.

Weber, R. P. 1990. *Basic Content Analysis*. 2nd edition. *Quantitative Applications in the Social Sciences Series*. Newbury Park, CA: Sage Publications.

Witten, I. H., and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd edition. San Francisco, CA: Morgan Kaufmann.

Yule, G. U. 1938. On sentence length as a statistical characteristic of style in prose with application to two cases of disputed authorship. *Biometrika* 30: 363–390.

Zhang, Q., M. Y. Hu, E. Patuwo, and D. C. Indro. 1999. Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European Journal of Operational Research* 116: 16–32.

Zhou, L., D. Twitchell, T. Qin, J. Burgoon, and J. Nunamaker. 2002. An exploratory study into deception detection in text-based computer-mediated communication. In *Proceedings of the 36th Hawaii International Conference on System Sciences*. Big Island, HI: Hawaii International Conference on System Sciences.

Zipf, G. 1929. Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology* 40: 1–95.

———. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.