Decision Support Systems 50 (2011) 595–601

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

A computational model for financial reporting fraud detection

Fletcher H. Glancy [a,*], Surya B. Yadav [b]

[a] Lindenwood University, 209 S. Kingshighway, Saint Charles, MO 63301, United States
[b] Dept. of Information Systems and Quantitative Sciences (ISQS), Rawls College of Business, Texas Tech University, Lubbock, TX 79409, United States

ARTICLE INFO

Available online 18 August 2010

Keywords:
Fraud
Financial reporting
Text mining
Quantitative model

ABSTRACT

A computational fraud detection model (CFDM) was proposed for detecting fraud in financial reporting. CFDM uses a quantitative approach on textual data. It incorporates techniques that use essentially all of information contained in the textual data for fraud detection. Extant work provides a foundation for detecting deception in high and low synchronicity computer-mediated communication (CMC). CFDM provides an analytical method that has the potential for automation. It was tested on the Management's Discussion and Analysis from 10-K filings and was able to distinguish fraudulent filings from non-fraudulent ones. CFDM can serve as a screening tool where deception is suspected.

1. Introduction

Corporate fraud has not been confined to the well-advertized cases of Enron, WorldCom, HealthSouth, etc. In all the AAERs we examined, the fraud detection was after years of abuse by senior management; and the Securities and Exchange Commission (SEC) did not detect it proactively. The continued pattern of fraud has shaken the confidence of the public in corporate America [28] both academics and auditing firms have been searching for ways to detect corporate fraud. While academic fraud research has examined many business areas [34], very little effort has been made to use quantitative approaches to examine textual data for automated financial reporting fraud detection. Phua et al. summarized the status of fraud research into four primary areas: internal, insurance, credit card, and telecommunications. In most of the internal fraud research, the object was to detect employee fraud or theft; financial reporting fraud involving senior management was not a major research focus. Most attempts to detect financial reporting fraud use financial ratios, applying various methodologies with varying results [6,28,30,31]. Phua et al. concluded that the use of unstructured data in fraud detection is essentially unexplored.

This paper proposes a quantitative model for detecting fraudulent financial reporting. The model detects the attempt to conceal information and/or present incorrect information in annual filings with the US Securities and Exchange Commission (SEC). The model uses essentially all of the information contained in a text document for fraud detection. A consistent and accurate screening tool would provide decision support for early detection of fraud; and hopefully, early detection will provide a deterrent to the commission of fraud.

In order to detect fraud, we must first define it. This would seem an easy task, but it is not always as straightforward as finding the dictionary definition [1]. We use the SEC's issuance of an Accounting and Auditing Enforcement Release (AAER) as a starting point for defining financial reporting fraud.

An AAER is an administrative proceeding or litigations release that entails an accounting or auditing related violation of the securities laws as enforced by the Securities and Exchange Commission (SEC). In the period from 2000 to 2008, the SEC issued 1700 AAERs. In the period from 2006 to 2008, they issued 555 AAERs. In this analysis, we use the term fraud when referring to an AAER to be a litigation release of an accounting or auditing violation where the SEC used the word fraud in describing the violation. We examined a sample of 74 AAERs from this period that charged companies with fraud. They showed that the average time between the identified initial fraud and the SEC filing charges was 7.26 years with a range of 3.6 to 11.6 years. The SEC charged that these companies committed fraud for an average of 4.1 years with a range of 1 to 12 years.

Churyk et al. [13] used qualitative content analysis of the required Management's Discussion and Analysis (MDA) part of the 10-K SEC filings to identify fraudulent filings. They were able to identify deceptive cues. SEC filings include several areas of text in addition to the MDA. All companies are required to explain anything in the operation that could have a significant impact on the future profitability of the company. Management's explanation is in the MDA; the accountant's explanation is in the notes to the consolidated financial statements. It has been proposed that if a company were to report information that would have potentially negative results on the company valuation, they would include it in the notes. The reason to include the information is that it is legally required, and failure to do so is a criminal act that could result in a long incarceration. Both fraud and failure to properly report are crimes; but reporting gives some protection to the accountants and auditors, which leaves only senior

* Corresponding author.
E-mail address: fglancy@lindenwood.edu (F.H. Glancy).

0167-9236/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2010.08.010

management at risk. The idea that information can be concealed in either the MDA or in the notes, as it can be said "in plain sight", leads to the problem statement and subsequent research questions.

1.1. Problem statement

The status of fraud detection from the analysis of corporate financial statements can be described as follows:

1. Fraud detection is after the fact and officially recognized only after the SEC issues an AAER.
2. There are few proven quantitative methods to detect a potentially fraudulent filing [28]. We did not find any quantitative method that examined the text of the filing.
3. A high level of senior management compensation comes from stock options. The individual manager has high potential return from financial fraud and a low potential for detection. This leads to the conclusion that corporate fraud is likely to continue at the senior executive level [22].

There is little question that the incentive to commit fraud at the senior executive level exists in spite of the laws passed since the Enron failure [22]. From Agency Theory, the most common executive compensation methods promote fraud at the senior executive level. The options backdating scandals that came to light in 2006 re-enforces that the incentive for executive fraud continues to exist [35]. It has been proposed that the "New Economy" makes fraud a normal circumstance because of the requirement to predict and deliver continually increasing earnings [39]. With the apparent widespread financial reporting fraud, there is a need for improved detection.

There has been minimal use of data mining in the investigation of fraud in financial reporting [29]. Text mining is a subset of data mining that has been used infrequently with varying results [5,12]. The Churyk et al. [13] finding that a qualitative methodology can provide an indication of fraud begs the question of creating a quantitative analytical model for text that has the potential for automation. Text mining is an obvious possibility. "(Text mining) is useful because it provides an efficient, quantitative representation of each document." [2]. There is a need to put text mining into context to produce positive results [18]; this requires that supervised text mining is used for the initial model. This indicates that there is a potential for its use in fraud detection, but that new techniques will be required. We know that there are indications of fraud in the 10-K filings [37]. This is based on the requirement of all companies to explain anything in the operation that could have a significant impact on future profitability. This requirement creates a situation where, if the 10-K writers know of an instance or situation that can have an impact and do not report it, they knowingly deceive the reader [22]. This knowing deception has been shown to create internal conflict that manifests itself in several ways [23]. The 10-K writer has an internal conflict because the failure to report is a crime as is fraudulent reporting. However, reporting honestly can result in a dramatic decrease in stock price [39] and a personal loss of wealth as well as employment. Another possibility exists, to report the instance but to do so in a manner to reduce the apparent effect on the future business by use of affect modifiers and other linguistic methods. The concealing or covert revealing of information in financial reporting leads to our research issues.

1.2. Research issues

This article investigates the following issues:

1. Can the fraud be detected from the text of annual corporate SEC filings?
2. Can a quantitative and computational-based fraud detection model be developed that will provide a methodology for automating detection of potential fraud?

We propose an answer to these questions by creating a model for detecting financial reporting fraud in corporate annual SEC filings and testing it using additional SEC annual filings. The research approach used for this investigation was based on the Unified Research Methodology [4] and design science [25].

The paper is organized into the following six major sections. Section 2 reviews the financial reporting fraud literature, discusses the potential theoretical basis for fraud detection in text, and a tool used for pattern detection in text. In Section 3, we develop the computational model. In Section 4, we test the model. Section 5 presents the results of the testing and analyzes the results. In Section 6, we present the contributions of this work, and Section 7 presents our conclusions and potential future work.

2. Relevant literature

There are two separate research literature streams providing background and basis for this research; they are research into corporate financial reporting fraud and text-mining research. Both of these research areas are extensive, but there has been very little crossover research. In order to understand the basis for assuming it is possible to detect fraud from text, we examine the body of research into deception. First, we look at the theoretical basis for deception detection, fraud detection, and then at text mining.

2.1. Deception detection

McCornack [32] created the Information Manipulation Theory (IMT) using Grice's cooperative principle of communication and the maxims of expected quality, quantity, relevance, and manner (or clarity). The deception occurs when there is a covert violation of one of the maxims. IMT states that the four principles are independent, and violation of a single principle defines a deceptive communication. IMT was tested and the replication of McCornack's work showed that the maxims were not independent and that any deception would violate the quality maxim [26]. Intuitively, this is logical because violations of the quality maxim involve distortions or fabrications of information.

Burgoon and Buller [9] proposed the Interpersonal Deception Theory (IDT) as an explanation of how deceivers react when interacting with a person that is the target of deception. They define deception as "a deliberate act perpetrated by a sender to engender in a receiver beliefs contrary to what the sender believes is true to put the receiver at a disadvantage." They later refined deception as "a message knowingly transmitted by a sender to foster a false belief or conclusion by the receiver" [8]. They found that "deceivers were more uncertain, and vague, more non-immediate and reticent, showed more negative affect, displayed more arousal and non-composure, and generally made a poorer impression than truthtellers. Their behaviors also connoted greater formality and submissiveness." They found that deceptive interactions are dynamic; they change over time as the deceiver attempts to manage image and the victim's interpretation. The presence of deception and degree of suspicion affected the target individual's behavior and the degree they mirrored the deceiver's behavior. Empirical tests have confirmed the dynamics of interaction and the manifestations of deception. A main principle of IDT is that the face-to-face communication is interactive, and the deceiver alters the deception based on the feedback received from the receiver [10]. Burgoon et al. show empirically that interaction is a primary contributor to IDT; and without interaction between the deceiver and their subject, the deception is very difficult to detect. We find that other researchers use IDT as a theoretical basis for deception detection in other media.

George et al. [20] used IDT in combination with Media Richness Theory (MRT) as the theoretical basis for examining if alerting the receiver of the possibility of deception in CMC would increase the

ability to detect the deception. The MRT has often been used to explain differences between media and the ability of some media to reduce equivocality and uncertainty in communications between managers and subordinates. The fit of the media to the message is necessary for successful communication. The dimensions of media are speed of feedback, language variety, personal focus, and social cues. IDT and MRT were combined with Conversation Theory to examine deceptive behavior in instant messaging [41]. Conversation Theory defines social systems as language oriented and symbolic. The response to one person's communication is based on the interpretation of their behavior, and meanings are reached through negotiated conversations. Zhou [41] placed instant messaging in the domain of synchronous communication, from which one could conclude that interaction is relative.

In testing asynchronous CMC, Hancock et al. [23] used both Communication Accommodation Theory (CMT) and Interpersonal Deception Theory as a basis for their empirical work. CMT suggests that when participants are trying to persuade or gain the approval of their partner, they tend to match a variety of behaviors, including accent, loudness, vocabulary, grammar, and gestures. IDT models interactive and ongoing transactions. The CMC test used a stylometric analysis, and involved partnered asynchronous communication. Stylometric analysis looks at the elements of a text, words and punctuation, without regard for context.

Carlson et al. [11] proposed a model of deception and detection for CMC that drew on IDT, MRT and Social Presence Theory. Social presence is the subjective and cognitive synthesis of the factors that reflect the intimacy of the communication medium. It is the sense of being with the other person and includes all of the linguistic and paralinguistic cues, including engagement, both psychologically and behaviorally. The authors describe multiple constructs for deception research: synchronicity, symbol variety, cue multiplicity, tailorability, reprocessability, and rehearsability. Using these constructs they create a propositional framework for the future testing. The authors state that the framework "serves to identify several areas to create tests and make knowledge claims." It was suggested that the framework could provide direction for automation of deception detection. The framework provides a solution for applying IDT to media of different synchronicity; the constructs are continuous not binomial.

Deception detection in low synchronicity CMC was tested using IDT as the theoretical basis through analysis of the linguistics based cues between paired individuals [42]. This analysis used 27 linguistic cues and clustered them into nine linguistic constructs. The results supported that deceptive senders create longer messages that tend to be informal uncertain, non-immediate, less complex and diverse than those of truth tellers. They also concluded that deceivers demonstrate these qualities in their behavior and showed more negative than positive affect.

Previous research has shown that many of the same cues that identify deceptive communication in face-to-face communications are present in CMC regardless of synchronicity. The text prepared for regulatory agencies has not been investigated for deceptive communication cues. Following IDT [8] and Carlson et al. [11], we suspect that deception cues are present if the person preparing text of the 10-K has knowledge of the deception and is motivated by self interest. If we are able to detect deception at the level of criminal deception or fraud, we can logically guess that the writer had knowledge of the deception. We next look at financial reporting fraud detection literature.

2.2. Fraud detection

Most of the fraud detection research utilizing data mining techniques has focused on structured data using quantitative methods [29]. The major categories of fraud previously investigated have been classified into four areas: internal, insurance, credit, and telecom [34]. In the last three areas, the emphasis is early identification of external attempts to commit fraud against a company [17]. The internal fraud research focuses primarily at detecting employee theft at a low level in the organization. Hake [22] argues that the most damaging fraud inflicted on an organization is at the senior executive level. The type of fraud committed at Enron and WorldCom resulted in the loss of all owner equity. He goes on to argue that compensation programs for senior executives utilizing stock options encourage this behavior; and Sarbanes–Oxley, while intended to prevent this behavior, does nothing to remove the incentive.

Prior quantitative research into financial reporting fraud concentrated on analysis using financial ratios as the variables [28]. Methods tested included cascaded logit and probit models [6,37], case base reasoning [40], decision trees, neural networks, and Bayesian belief networks [30]. These were not able to reliably predict fraud.

Few researchers have looked at the text in financial reports. Churyk et al. [13] used qualitative content analysis on the MDA of SEC 10-K filings. They were able to find differences between companies that restated earnings, as the AAER required, and the sample of companies that were not required to restate earnings. Text analysis was done on the MDA utilizing the software program Linguistic Inquiry and Word Count (LIWC). Pennebaker et al. [33] state that "In order to provide an efficient and effective method for studying the various emotional, cognitive, and structural components present in individuals' verbal and written speech samples, we originally developed a text analysis application called Linguistic Inquiry and Word Count, or LIWC." Most text analysis or context analysis traces the methodology to Glaser [21]. The method Glaser proposed and Pennebaker et al. used in the creation of LIWC requires coding of text. Some researchers argue that the qualitative data from text analysis can be converted into quantitative results [36], but this is not widely accepted as a quantitative methodology. There is, however, a proven method for analyzing text quantitatively based on term and document patterns.

2.3. Pattern detection in text

The relationships between words in documents form patterns. Detection of these patterns is the primary focus of text mining (TM) [2]. TM is a quantitative methodology that uses mathematical methods to describe the term-document matrix and increases the density through use of a singular value decomposition vector (SVD) [2,19]. TM is a subset of data mining that refers to techniques used to discover unknown information from natural language documents [24]. TM converts text to structured data. A primary method for dimension reduction, while retaining the maximum amount of information, is through creation of SVD [2]. This conversion is unidirectional. The results are no longer context specific; the results are statistics that give an indication of the significance of the text. The actual relationships of the context are still present. Current TM methods do not support recovery of context from the SVD, but the original documents are retained and the relationship between them can be seen. The process is not readily or easily reversible. The information that is contained in the term-document context is retained. See Albright [2] for a detailed discussion of the SVD. For example, an SVD with a dimension of 100 by 10 retains all of the original information, but is not easily interpretable into natural language. Text mining allows the terms of document sets to be independent variables and compared against dependent variables in supervised TM. In unsupervised TM, the methods are primarily exploratory. However, the patterns resulting from exploratory TM clustering are observable and useable as input for additional data mining methods.

Text mining has been used for fraud detection. The primary types of fraud were insurance and credit card [15]. This example used supervised TM, which creates a specific data set for training the TM for fraud detection in a new data set. Supervised text mining requires a target variable to create the transformed variables from the text. In

order to use TM in the detection of fraud, we create a computational model that uses TM in a manner that is replicable and reusable. In the next section, we develop this computational model.

3. Computational model of fraud detection

We have seen in the previous section that detection of deceit is possible in CMC that has low synchronicity. We propose to show that it is possible to detect deceit, fraud in the case of SEC filings, in documents that are essentially anonymous, in that neither writer nor reader are specifically identified or known to each other. This process is based on several concepts with a theoretical foundation in Interpersonal Deception Theory and Media Richness Theory. The concepts are maximal information usage, writer's knowledge, and informational cues in the development of CFDM. The process retains maximal information and uses essentially all of it in processing the documents. We conceptualize that the deceit is detectable because the writer has knowledge of the deceit. And as the writer has knowledge, she/he provides detectable cues of the deceit.

The 10-K filings are prepared by the filing company and directed to an unspecified person who has some interest in reading them. We propose that the same mechanisms in effect in person-to-person deceit are in effect in the preparation of required SEC filings. We develop the computational fraud detection model (CFDM) using SAS® Enterprise Miner™ (EM) as an automation tool to develop the model. The computational model is designed to be reusable in different domains and fully supported by EM. The model is shown in Fig. 1 in the form of a process flow diagram.

The model involves the following process steps:

1. Select the target companies and target document.
2. Extract the target text from the document.
3. Prepare the target text for importing into EM.
4. Import all of the target text documents into EM creating a database.
5. Stem terms to reduce the dimensions and to tag parts of speech.
6. Create SVDs to further reduce the document dimensions.
7. Cluster the document set.
8. Review and evaluate the clustering results.
9. Modify the SVD and clustering algorithm as necessary.

Following the process flow for this computational model will allow application of the model in different domains.

To test this model, we apply the model to a database created from the 10-Ks filed by a selection of companies that have been accused of fraud by the SEC. We matched these companies with companies in the same industry that have not been charged with financial reporting fraud. We discuss the CFDM process as we apply it to the test database in the following sections.

4. CFDM implementation and testing

We have implemented CFDM using EM as the automation tool. We discuss the CFDM steps in detail below.

4.1. Sample company selection

Companies were selected from those SEC issued an AAER citing the company for fraud during the years of 2006 to 2008. The four digit SIC code was used to define the industry. The study included no more than two companies with the same SIC code. Each company that had been accused of fraud by the SEC was matched with a company in the same SIC code and of approximately the same size that had not been charged with fraud; and therefore, had not submitted an amended 10-K because of an AAER. Because innocence is considerably more difficult to prove than guilt, we set a high standard for the matching company. The primary screening was for companies that had not amended their 10-K in the last ten years. The only exceptions were for two companies that had not amended a 10-K in the last eight years and their amendment

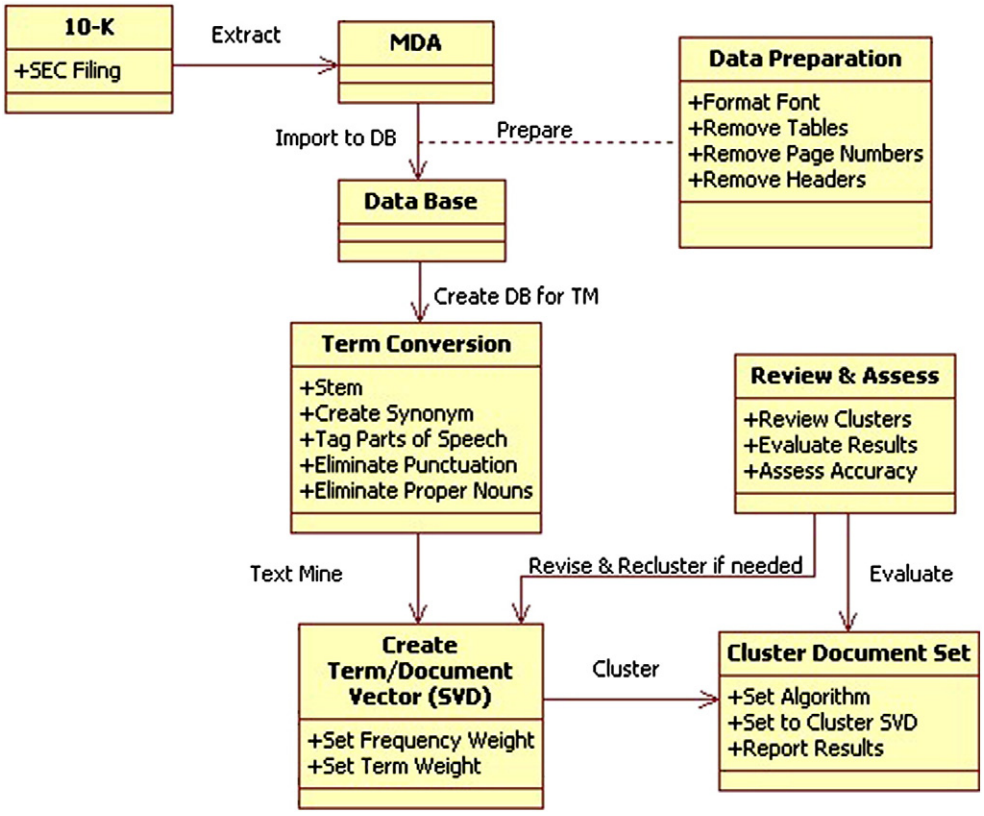

Fig. 1. Computational model for textual fraud detection.

was for peripheral data. This exception was made due to the difficulty of finding companies in certain industries that had not filed amended annual reports.

The target document is the10-K. The preferred 10-K for testing the model was at least one year prior to the latest period cited in the AAER. For example, if the AAER complaint accused the company of committing fraud for period from 2002 to 2005, the 10-K for 2004 was selected for the analysis. If the selected 10-K mentions that the SEC is investigating the company, then the 10-K for the prior year was selected. Because the SEC investigation often takes an extended period, there were several cases where we had to go back more than one year. The matching company 10-K was from the same time period. This was to reduce any potential bias caused by changing accounting rules, legal requirements, or other external influences.

4.2. Target text selection

There are two major blocks of text in a 10-K, the MDA and the Notes to the Financial Statement. An argument can be made for using either or both. Senior management assists in the preparation of the MDA and in some cases may write portions. From IDT, if fraud exists and the writer knows that it exists and has an incentive to deceive the reader, the MDA should provide clues of deception or fraud even with the low synchronicity of the MDA. The accountants that prepared the financial statements also prepared the notes. When financial reporting fraud exists, the notes writer will have knowledge and an interest in concealing it. We decided against using both because of the large size of the combined sections. Since we are attempting to detect financial reporting fraud that would involve and likely originate from senior management, we chose to use the MDA as the target text.

4.3. Text preparation

The SEC required format for 10-K's allows for latitude in the submissions. The typical formats are html, text, and the text source for html. For maximum accuracy in text mining, the data format needs to be consistent across documents. The MDA, Sections 7 and 7(a) of the 10-K and Section 6 of the 10-KSB, were copied from the 10-K to a word processing program; MS Word was used in this analysis. All documents were re-formatted with sans serif font; Arial was used. A preliminary test showed that serif fonts are subject to incorrect interpretation by the text-mining program, SAS Enterprise Miner 4.3 (EM). All tables were deleted from the documents. The text-mining program will not read documents that contain tables. Where the tables were used for document format and only contained text, the text was copied to the body of the document and the table deleted. The documents were saved in Rich Text Format (RTF) because the text-mining program was not compatible with the current version of MS Word (2007). These were no other manipulations of the text. The goal was to retain the maximum amount of textual data in the MDAs.

4.4. Document importing

All documents were imported into a single SAS data set with the full document text retained in an external HTML document. Retaining the entire document externally allows an unlimited amount of text to be used for a single document and included in the analysis. A binary target was added to the data set to indicate the presence or absence of an SEC issued AAER.

4.5. Term conversion

SAS EM as used in this analysis utilizes synonyms to create equal meaning words, utilizes parts of speech to create separate words, eliminates punctuation, and eliminates common words that could impede the text-mining process. The words or terms were stemmed, but the part of speech was retained. This allowed treating the same word used as different parts of speech as separate words. It increases the dimensions of the document term matrix, but can improve the results. If part of speech is not retained and the words are stemmed, terms such as *bank* and *banking* are treated as a single term, so a financial institution, a bank, is equated to banking an airplane. Punctuation was not included in this analysis. A Standard English stop list that eliminates common terms and a synonym list were used with addition of the terms management, discussion, and analysis to the stop list. Since the document set was the Management's Discussion and Analysis of the annual report these terms were in all documents and added no information.

4.6. SVD creation and clustering

SVDs were created using the full term data set at a low resolution. Log frequency weighting and information gain term weighting were used. A detailed explanation of frequency and term weighting is beyond the scope and length limitation of this paper. Both expectation maximization and hierarchical clustering were done. The last two parts of the CFDM process, (8) results review and evaluation and (9) SVD Modification and Iteration are discussed in Section 5.

5. Results and discussion

We continue the description of the CFDM through the results, evaluation, and testing.

5.1. CFDM results review and evaluation

Supervised expectation maximization clustering and hierarchical clustering were performed on the SVDs created with log frequency weights and information gain term weights. The expectation maximization clustering was unstable because of the presence of local minima and the selection of the starting point for the partitive processing determined the endpoint [7,27,38]. The hierarchical clustering was stable with the agglomerative clustering reaching the same end point for all clustering trials. Repeated reclustering tested both methods.

The documents clustered into two stable clusters. In all cases, the CFDM separated the documents into two clusters regardless of the allowed maximum number of clusters; five, ten, and forty clusters were allowed on consecutive trials. The clustering resulted in no false negatives and three false positives in sixty-nine sets of documents. The results of the hierarchical clustering were evaluated using the sign test [3,16]. The null hypothesis is that if there is no ability of the CFDM to discriminate between fraud and no fraud; the probability distribution in each cluster is 0.5.

As shown in Table 1, the clustering results were able to identify potential fraud in the 10-Ks at a level of significance greater than 0.01 (the calculated p-value was 1.2×10^{-14}). The statistical power was approximately 90% [14]. Table 2 gives the highest weighted terms that define the clusters.

Table 1
Text mining hierarchical clustering results and prediction at 0.01 significance level.

CFDM	Total	Text mining results			Predicted at 0.01 level	
		Correctly identified	Incorrectly identified	p-value	Upper	Lower
MDA	69	66	3	1.2×10^{-14}	45	24

5.2. CFDM testing

A validation test of the CFDM was performed to test for the ability to discriminate MDAs of companies that had committed fraud and for the ability to discriminate MDAs of companies that had not committed fraud. The CFDM was first tested on a new sample of ten individual companies that met the criteria for the selection of the original companies that received AAERs accusing them of fraud. The MDA preparation was the same as described earlier in order to preserve the maximum data in the text. The MDAs were tested individually using the CFDM. Of the ten fraudulent MDAs, the CFDM clustered nine to be fraudulent. We examined the one false negative to see if additional information could be gained. This MDA was from the first year the SEC cited in the AAER, and the SEC said that the fraud started mid-year. The MDA from the following year was tested using the CFDM, and it clustered as fraudulent. We included all of the MDAs tested to avoid discarding relevant data. Evaluating the clustering with the sign test, the p-value for 10 of 11 MDAs clustering correctly is 0.0059 [16].

Because the CFDM development demonstrated a potential bias toward fraud detection due to only false positives and a concern that the effect size may be smaller for the detection of non-fraudulent MDAs, a new sample of twenty companies was chosen. The twenty non-fraudulent companies were selected using the same criteria for the development of the CFDM. None of these were in the original data set. The MDA preparation was consistent with the CFDM and the test of fraudulent companies. Four of the 20 non-fraudulent companies clustered as fraudulent. Using the sign test, the results from the twenty companies give a p-value of 0.0059 [16]. As shown in Table 3, testing of the CFDM demonstrates that it discriminates at a significant level for both fraudulent and non-fraudulent MDAs.

Table 3
CFDM testing results and prediction at 0.01 significance level.

Sample	Total	Text mining results			Predicted at 0.01 level	
		Correctly identified	Incorrectly identified	p-value	Upper	Lower
Fraudulent MDAs	11	10	1	0.0059	9	2
Non-fraudulent MDAs	20	16	4	0.0059	16	4

6. Research contribution

We presented a new and effective computational model to detect fraud using text-mining techniques. CFDM contrasts with current quantitative models that have been proposed for financial fraud detection. The other models use financial ratios as input, and CFDM uses the data contained in text as input. The CFDM demonstrated a significant ability to discriminate fraudulent companies from non-AAER companies using the MDA; and thus, it answered the first research issue.

The first research issue was: can the fraud be detected from the text of annual corporate SEC filings? Fraud was detected from the MDA text. The ability of the CFDM to discriminate fraudulent text implies that the writer of the MDA was under some kind of stress that affected the writing. It may be that the writer had knowledge or at least belief that the document represented financial reporting fraud. The stress may have had other causes. We suggest that the presence of deceit indicating cues is significant, and it could be inferred that the writer believed that the MDA was representing deception. CFDM detected the cues in a low presence, high rehearsability, and low synchronicity media. Low presence is the lack of interaction. The text is presented without any indication of the writer's identity. High rehearsability is that without interaction, the writer has the time and ability to edit the text. The text of a filing with the SEC has essentially no interaction between the writer and the reader. There is no synchronous interchange between them.

Table 2
Eight highest weighted terms representing clusters.

	Top eight terms
Cluster 1 No AAER	Affect, require, expense, rate, do, cost, year, and determine
Cluster 2 AAER	Represent, account, relate, continue, reduce, reflect, sell, and expect

The successful implementation of the CFDM also answers the second research issue: can a quantitative and computational-based fraud detection model be developed that will provide a methodology for automating detection of potential fraud? CFDM is a quantitative and computational-based fraud detection model that provides a method for automating detection of potential fraud. The main contribution of this research is a novel quantitative computational-based model that uses maximal information for detecting fraud in textual data.

In addition, we assert two other significant contributions:

1. The computational model is reusable and has the potential to detect fraud in domains other than financial reporting.
2. CFDM has the potential to serve as a filtering tool for regulators to focus their resources and subsequently increase the detection of financial reporting fraud.

The companies in this study were selected because the SEC issued an AAER. The reason for the AAER was not a selection criterion; therefore, there were wide varieties of violations of SEC financial reporting and auditing regulations. They covered a large range of financial reporting fraud that included round-about transactions, falsification of sales, showing sales to distributors without recording the return, bribing foreign officials, expensing vacation trips for foreign officials and customers, etc. While there was a wide variety in the type of offense, they still clustered as committing fraud.

7. Conclusion and future research

The CFDM demonstrates that it is possible to detect financial reporting fraud from the text of annual filings with the Security and Exchange Commission. The model is generalizable because it specifies automatable steps that can be adapted to other domains and genres. A potential application for CFDM is to screen companies for investigation of potential fraud by the SEC. This would allow effective use of SEC resources. Additional potential applications include investor analysis, e-mail spam detection, and business intelligence validation.

This work can be criticized for limiting the training data set to sixty-nine companies. While the statistical power is over 90% for this sample size, further confirmation of the discriminatory power of the CFDM by increasing the sample size would be an area for future research. The domain of this work is limited to MDA in 10-K filings. We do not argue that it is directly generalizable to other text submissions, but that other domains are a fertile area for research using CFDM. The notes to the financial statements in the 10-K are a possible area to extend CFDM without changing the domain.

The model opens several additional research areas: (1) deception detection in e-mail and other computer-mediated communication; (2) deception in business-to-consumer websites and in consumer-to-consumer websites; (3) increasing the understanding of the mechanisms present in asynchronous text deception. Further work analyzing the singular value decomposition and the respective document and text vectors may lead to improved understanding of

the mechanisms present in deceptive communications. It would be interesting to see if the CFDM has the potential to be a tool for decision support through evaluation of the degree of the veracity in unstructured data provided by a business intelligence system.

References

[1] R.N. Aditya, The psychology of deception in marketing: a conceptual framework for research and practice, Psychology and Marketing 18 (7) (2001).
[2] R. Albright, Taming Text with the SVD, SAS Institute White Paper, Retrieved from www.sas.com/apps/whitepapers/whitepaper.jsp?code=SDM5 2004.
[3] J. Arbuthnott, An Argument for Divine Providence, taken from the constant Regularity observ'd in the Births of both Sexes, Philosophical Transactions 27 (1710).
[4] D. Baldwin, S.B. Yadav, The process of research investigaions in artificial intelligence — an unified view, IEEE Transactions on Systems, Man, and Cybernetics 25 (5) (1995).
[5] B. Ballou, J.M. Mueller, Helecom communications: considering fraud risk on an engagement before and after analyzing a key business process, Issues in Accounting Education 20 (1) (2005).
[6] M.D. Beneish, The detection of earnings manipulations, Financial Analysts Journal 55 (5) (1999).
[7] S. Borman, The Expectation Maximization Algorithm: A short tutorial, Retrieved from http://www.seanborman.com/publications/EM_algorithm.pdf 2009.
[8] D.B. Buller, J.K. Burgoon, Interpersonal detection theory, Communication Theory 6 (3) (1996).
[9] J.K. Burgoon, D.B. Buller, Interpersonal deception: III. Effects of deceit on perceived communication and nonverbal behavior dynamics, Journal of Nonverbal Behavior 18 (2) (1994).
[10] J.K. Burgoon, D.B. Buller, A.S. Ebesu, P. Rockwell, Interpersonal deception: V. Accuracy in deception detection, Communications Monographs 61 (1994).
[11] J.R. Carlson, J.F. George, J.K. Burgoon, M. Adkins, C.H. White, Deception in computer-mediated communication, Group Decision and Negotiation 13 (1) (2004).
[12] H. Chen, W. Chung, J.J. Xu, G. Wang, Y. Qin, M. Chau, Crime Data Mining: A General Framework and Some Examples, Computer 37 (4) (2004).
[13] N.T. Churyk, C.C. Lee, D.B. Clinton, Early Detection of Fraud: Evidence from Restatements, in: V. Arnold (Ed.), Advances is Accounting Behavioral Research Vol. 12, JAI Press, Bingley, UK, 2009.
[14] J. Cohen, Statisical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.
[15] R.S. Collica, CRM Segmentation and Clustering using SAS Enterprise Miner, SAS Institute Inc. Cary, NC, 2007.
[16] W.J. Conover, Practical Nonparametric Statistics, Third EditionJohn Wiley & Sons, Inc, New York, 1999.
[17] N. Conz, Mining Your Own Business — data mining technologies can help insurers access and leverage the institutional knowledge vital to fraud mitigation efforts that is locked inside their current and historical claims data, Insurance and Technology 32 (10) (2007).
[18] M. Ellingsworth, Mining Tools Put Text into Context, Optimize (2003) October 1.
[19] J. Gao, J. Zhang, Clustering SVD strategies in latent semantic indexing, Information Processing and Management 41 (2005).
[20] J.F. George, K. Marett, P.A. Tilley, The Effects of warnings, computer-based media, and probing activity on successful lie detection, IEE Transactions on Professional Communication 51 (1) (2008).
[21] B.G. Glaser, The constant comparative method of qualitative analysis, Social Problems 12 (4) (1965).
[22] E.R. Hake, Financial illusion: accounting for profits in an enron world, Journal of Economic Issues 39 (3) (2005).
[23] J.T. Hancock, L.E. Curry, S. Goorha, M. Woodworth, On lying and being lied to: a linguistic analysis of deception in computer-mediated communication, Discourse Processes 45 (1) (2008).
[24] M. Hearst, What is Text Mining? Retrieved from http://www.ischool.berkley.edu/~hearst/text_mining.html 2003.
[25] A.R. Hevner, S.T. March, J. Park, S. Ram, Design science in information systems research, MIS Quarterly 28 (1) (2004).
[26] S. Jacobs, E.J. Dawson, D. Brashers, Information manipulation theory: a replication and assessment, Communications Monographs 63 (1996).
[27] S.S. Kamaruddin, A.R. Hamdan, A.A. Bakar, Text mining for deviation detection in financial statements, Proceedings of the International Conference on Electrical Engineering and Informatics, Institut Teknologi Bandung, Indonesia, 2007, June 17–19.
[28] K.A. Kaminski, T.S. Wetzel, L. Guan, Can financial ratios detect fraudulent financial reporting? Managerial Auditing Journal 19 (1) (2004).
[29] E. Kirkos and Y. Manolpoulos, Data Mining in Finance and Accounting: A review of Current Research Trends, Paper presented at the Internation Conference on Enterprise Systems and Accounting (Thessaloniki, Greece 2004).
[30] E. Kirkos, C. Spathis, Y. Manolopulos, Data mining techniques for the detection of fraudulent financial statements, Expert Systems With Applications 32 (2007).
[31] S. Kotsiantis, E. Koumanakos, D. Tzelepis, V. Tampakas, Forecasting fraudulent financial statements using data mining, International Journal of Computational Intelligence 3 (2) (2006).
[32] S.A. McCornack, Information manipulation theory, Communications Monographs 59 (1) (1992).
[33] J.W. Pennebaker, C.K. Chung, M. Ireland, A. Gonzales, R.J. Booth, The LIWC Application, Retrieved from http://www.liwc.net/liwcdescription.php 2009.
[34] C. Phua, V. Lee, K. Smith, R. Gayler, A Comprehensive Survey of Data Mining-based Fraud Detection Research, Available at http://www.bsys.monash.edu.au/people/cphua/ (2005).
[35] M.A. Siegel, Options backdating, The CPA Journal 77 (19) (2007).
[36] K.J. Srnka, S.T. Koeszegi, From words to numbers: how to transform qualitative data into meaningful quantitative results, Schmalenbach Business Review 59 (2007).
[37] S.L. Summers, J.T. Sweeney, Fraudulently misstated financial statements and insider trading: an empirical analysis, The Accounting Review 73 (1) (1998).
[38] B.G. Tabachnick, L.S. Fidell, Using Multivariate Statistics, Fifth ed.Pearson, Boston MA, 2007.
[39] R. Tillman and M. Indergaard, Pump and Dump: Corporate Corruption in the New Economy, Paper presented at the American Sociological Association, (Annual Meeting 2003).
[40] R. Wheeler, S. Aitken, Multiple algorithms for fraud detection, Knowledge-Based Systems 13 (2000).
[41] L. Zhou, An empirical investigation of deception behavior in instant messaging, IEEE Transactions on Professional Communication 48 (2) (2005).
[42] L. Zhou, J.K. Burgoon, J.F. Nunamaker, D. Twitchell, Automating linguistics-based cues for detecting deception in text-based asynchronous computer mediated communication, Group Decision and Negotiation 13 (1) (2004).

Fletcher Glancy is an Assistant Professor of MIS at the School of Business and Entrepreneurship, Lindenwood University. He received his Bachelors of Science degree in mechanical engineering from Missouri S & T in 1970, his MBA from Texas Tech University in 2006, and his Ph.D. from the Rawls College of Business, Texas Tech University in 2010. He has 35 years of industry experience. His areas of interest include: business intelligence, text and data mining, linguistics, theory development, and analytical methodology.

Surya B Yadav is the James & Elizabeth Sowell Professor of Telecom Technology in Rawls College of Business, Texas Tech University, Lubbock, Texas. He received his Bachelors of Science degree in electrical engineering from Banaras University in 1972, the M.Tech. degree from IIT Kanpur, India in 1974, and the Ph.D. degree in business information systems from Georgia State University, Atlanta in 1981. He has published in several journals including Communications of the ACM, IEEE Transactions on Software Engineering, IEEE Transactions on Systems, Man, and Cybernetics, Journal of Management Information Systems, Decision Support Systems, and Journal of Intelligent Information Systems. His research areas include intelligent information retrieval systems, text mining, and system security.