View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by AIS Electronic Library (AISeL)

Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2014 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

2014

THE DETECTION OF FRAUDULENT FINANCIAL STATEMENTS: AN INTEGRATED LANGUAGE MODEL

Wei Dong
USTC-CityU Joint Advanced Research Centre, weidong1@mail.ustc.edu.cn

Stephen Shaoyi Liao
City University of Hong Kong, issliao@cityu.edu.hk

Bing Fang
Shanghai University, melodyfang@shu.edu.cn

Xian Cheng
USTC-CityU Joint Advanced Research Centre, chengcx@mail.ustc.edu.cn

Zhu Chen
City University of Hong Kong, zhuchen1025@gmail.com

See next page for additional authors

Follow this and additional works at: http://aisel.aisnet.org/pacis2014

Recommended Citation
Dong, Wei; Liao, Stephen Shaoyi; Fang, Bing; Cheng, Xian; Chen, Zhu; and Fan, Wenjie, "THE DETECTION OF FRAUDULENT FINANCIAL STATEMENTS: AN INTEGRATED LANGUAGE MODEL" (2014). *PACIS 2014 Proceedings*. 383.
http://aisel.aisnet.org/pacis2014/383

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2014 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Authors
Wei Dong, Stephen Shaoyi Liao, Bing Fang, Xian Cheng, Zhu Chen, and Wenjie Fan

This conference proceeding is available at AIS Electronic Library (AISeL): http://aisel.aisnet.org/pacis2014/383

THE DETECTION OF FRAUDULENT FINANCIAL STATEMENTS: AN INTEGRATED LANGUAGE MODEL APPROACH

Wei Dong, College of Business, USTC-CityU Joint Advanced Research Centre, University of Science and Technology of China, City University of Hong Kong, Hong Kong, weidong1@mail.ustc.edu.cn

Shaoyi Liao, College of Business, City University of Hong Kong, Hong Kong, issliao@cityu.edu.hk

Bing Fang, School of Management, Shanghai University, Shanghai, China, melodyfang@shu.edu.cn

Xian Cheng, College of Business, USTC-CityU Joint Advanced Research Centre, University of Science and Technology of China, City University of Hong Kong, Hong Kong, chengcx@mail.ustc.edu.cn

Chen Zhu, College of Business, City University of Hong Kong, Hong Kong, zhuchen1025@gmail.com

Wenjie Fan, College of Business, City University of Hong Kong, Hong Kong, wj.fan@cityu.edu.hk

Abstract

Among the growing number of Chinese companies that went public overseas, many have been detected and alleged as conducting financial fraud by market research firms or U.S. Securities and Exchange Commission (SEC). Then investors lost money and even confidence to all overseas-listed Chinese companies. Likewise, these companies suffered serious stock sank or were even delisted from the stock exchange. Conventional auditing practices failed in these cases when misleading financial reports presented. This is partly because existing auditing practices and academic researches primarily focus on statistical analysis of structured financial ratios and market activity data in auditing process, while ignoring large amount of textual information about those companies in financial statements. In this paper, we build integrated language model, which combines statistical language model (SLM) and latent semantic analysis (LSA), to detect the strategic use of deceptive language in financial statements. By integrating SLM with LSA framework, the integrated model not only overcomes SLM's inability to capture long-span information, but also extracts the semantic patterns which distinguish fraudulent financial statements from non-fraudulent ones. Four different modes of the integrated model are also studied and compared. With application to assess fraud risk in overseas-listed Chinese companies, the integrated model shows high accuracy to flag fraudulent financial statements.

Keywords: Financial fraud, Fraud detection, Statistical language model, Latent semantic analysis, Integrated model.

1. INTRODUCTION

Financial fraud is a serious problem worldwide. Wang et al. (2006) defined financial fraud as "a deliberate act that is contrary to law, rule, or policy with intent to obtain unauthorized financial benefit". Vivid cases of high-profile frauds include Lehman Brothers, WorldCom, and Enron (Bankruptcydata.com 2012), which have enormously influenced the global economy and the stock markets. With the undergoing of financial transformation in China, many leading mainland-based companies choose to list themselves overseas to gain access to investor capital. A set of stocks of these companies are called China Concept Stock[1]. Since the start of 2010, many Chinese companies that went public in North American capital market were shorted by a number of market research firms (e.g., Muddy Waters Research and Citron Research). Companies that have been alleged for fraud activities, including ONP (Orient Paper, Inc.), RINO (RINO International Corp.), DGW (Duoyuan Global Water, Inc.), and CCME (China Media Express Holdings, Inc.), suffered serious stock sank, and then been investigated by U.S. Securities and Exchange Commission (SEC), and finally they were even delisted from stock exchange. Many investors lost money and even confidence to all overseas-listed Chinese companies. In case of these great lost, there is an urgent need to detect and identify financial fraud, which is instrumental to ensure a fair, open, and transparent financial market.

Though the general guidelines from anti-fraud laws facilitate auditors to detect fraudulent financial statements, it remains as a different task using normal auditing procedures (Coderre 1999). Conventional auditing practices primarily focus on statistical analysis of structured financial ratios and market activity data in financial statements (Cecchini et al. 2010a; Ravisankar et al. 2011). Since the financial statements are well-planned and ex-anti prepared reports, the management has enough time and abilities to conceal the true financial condition by accounting shenanigans. What's more, under traditional corporate audit mechanism, top manager of a firm who is familiar with audit practice is growing wiser and wiser to circumvent different audit mechanism. Hence, conventional fraud detection methods always work not so well in nowadays circumstances.

Recently, researchers have begun to look at textual data in financial statement, i.e., Management's Discussion and Analysis (MD&A) section, to help better understand the dynamics of financial fraud in this previously untapped information (Cecchini et al. 2010b; Glancy & Yadav 2011; Humpherys et al. 2011). These researchers attempted to extract linguistic cues from textual information, and viewed the detection of fraudulent financial statements as a text classification problem. We think that research on applying text mining techniques to detect fraudulent financial statements is still at its infancy. There are problems with these linguistic cues-based fraud detection approaches (Zhou et al. 2008). First, these researches depended highly on the extraction of a predefined set of cues, which may be impeded by ineffective cue identification. Second, extracting cues from text is labor intensive and may involve ad hoc decisions. Not all of the cues can be extracted automatically due to the complexity and ambiguity of natural language. Third, the dependency relationships between words are not considered.

To address the above problems, we propose an integrated language model for detection of fraudulent financial statements. First, statistical language model (SLM) is used to estimate the occurrence probability of a span of text, which can be adapted to detect the strategic use of deceptive language in financial statements. This approach is significantly superior to existing linguistic cues based methods from at least two perspectives. One is that we do not need to extract a predefined set of cues in advance, which is time-consuming and labor-intensive. The other is that SLM can essentially model the dependency relationships between words in natural language, which is of great importance but is ignored in previous research. However, SLM is self-limited in detecting fraudulent financial statement because of its inability to capture long-span information. To overcome this insufficiency, latent semantic analysis (LSA) framework, a method to calculate document similarity, is integrated with SLM. The LSA is also able to extract the semantic patterns which distinguish fraudulent financial statements from non-fraudulent ones.

[1] Definition from Wikipedia (http://en.wikipedia.org/wiki/China_Concepts_Stock).

This research contributes to the problem of financial fraud detection from both theoretical and empirical perspective. From a theoretical perspective, we deal with four integration modes of two language models, i.e., SLM and LSA model. The integrated model not only overcomes the inability of SLM to capture long-span information, but also extract the semantic patterns from textual information. From an empirical perspective, we verify that language model can serve as an effective approach in detection of fraudulent financial fraud. The integrated model shows high accuracy to flag fraudulent financial statements. This research will benefit financial governors and auditors in detecting fraud and protecting the public's investments.

This paper unfolds as follows. Section 2 reviews the extant literature and defines the research gap. Section 3 presents the theoretical background and develops an integrated model. Section 4 discusses the sample selection and data collection process. Section 5 deals with the empirical application, including data preparation and data analysis procedures. Conclusions come in the last section.

2. LITERATURE REVIEW

As financial fraud detection is a classical research topic, there are plenty of researchers coming from various disciplines. In order to have a general view, we review about 30 papers range from 1990 to 2012, and generalize different financial and nonfinancial measures together with the methodologies used to predict financial fraud.

In terms of measure selection, a common thread in prior literature is to develop a checklist for auditors in their auditing practice. The items in checklist are called red flags, which is discussed in SAS No. 88. The use of red flags in assessing the risk of management fraud was pioneered by Albrecht & Romney (1980). Loebbecke & Willingham (1988) developed an L/W model containing three conditions under which fraudulent financial reporting might be perpetrated. The three conditions, i.e., conditions (C), motivation (M), and attitude (A), contained 46 red flags. However, Cecchini et al. (2010a) mentioned that previous literatures figured out that there are some problems with red flag checklist as a fraud detection mechanism.

Another popular thread is employing quantitative financial ratios to detect potential financial fraud. Considering the financial ratios used, prior studies unvaryingly employed data taken from annual financial reports. Most studies used 8 to 10 financial ratios (Beneish 1999; Green & Choi 1997; Kirkos et al. 2007; Persons 2011; Spathis 2002). Other researchers used large feature sets, almost more than 20 ratios (Kaminski et al. 2004). Two kinds of financial ratios appeared in existing literatures. Except for the ratios of a certain fiscal year, ratio trends (or year-over-year changes of ratios) are widely used (Cecchini et al. 2010a; Lin et al. 2003; Liou 2008).

Some researchers combined both quantitative variables (including financial ratios) and qualitative variables (including the red flags) together into an combined model. Fanning & Cogger (1998) screened out 20 variables, containing financial ratios and nonfinancial indicators from following aspects such as corporate governance, auditor, agency issues, subsidiaries, capital structure, operating results, personnel, litigation, accounting choices, financial statement accounts and ratios, trend analysis. Summers & Sweeney (1998) added a specific risk factor, insider trading, which was neither identified as a red flag in SAS No.82 nor in prior empirical research. Gaganis (2009) identified 28 financial ratios covering all aspects of firms' performances, such as liquidity, leverage, profitability, managerial activity and annual changes in basic accounts, together with six nonfinancial variables, i.e. financial distress, client litigation, audit firm, auditors switching, loss or profit in the year of audit opinion and whether a company is listed or unlisted. Abbasi et al. (2012) incorporated industry-level and organizational context-based features.

In order to handle these different measures, we summarize four types of methodologies used in literature (first four lines in Table 1). These researches always used just one method, and compared their results with previous researches. In contrast, some researchers employed more than two methods and made an overall comparison (Gaganis 2009; Kirkos et al. 2007; Kotsiantis et al. 2006; Liou 2008; Perols 2011).

Methodology	Specific technique	References	Total
Statistical model	Discriminant analysis	Kaminski et al. (2004)	1
	Logistic regression model	Bell et al. (2000); Dechow et al. (2011); Yuan et al. (2008); Liou (2008); Persons (2011); Spathis (2002); Summers & Sweeney (1998)	7
	Probit model	Beneish (1999)	1
Fuzzy set theory	Fuzzy set model	Deshmukh et al. (1997); Deshmukh & Talluru (1998)	2
Multi-criteria decision model	UTADIS and MHDIS model	Pasiouras et al. (2007); Spathis et al. (2002)	2
Data mining	Neural network	Fanning & Cogger (1998); Green & Choi (1997); Kirkos et al. (2007); Kotsiantis et al. (2006); Lin et al. (2003); Liou (2008)	6
	Decision tree	Bai et al. (2008); Kirkos et al. (2007); Kotsiantis et al. (2006); Liou (2008)	4
	Bayesian belief network	Kirkos et al. (2007); Kotsiantis et al. (2006)	2
	K-nearest neighbours	Kotsiantis et al. (2006)	1
	Support vector machine	Cecchini et al. (2010a); Kotsiantis et al. (2006)	2
Text mining	Linguistic cues based method	Cecchini et al. (2010b); Humpherys et al. (2011); Larcker & Zakolyukina (2012)	3
	A computational fraud detection model	Glancy & Yadav (2011)	1

Table 1. Methodologies used in prior fraudulent financial statement detection studies.

Although these prior measures and methodologies, especially the data mining methods, dominate the community of both practitioners and academic researchers, some problems are inevitable. For these researches, results varied among various methodologies. In addition, even though large amount of qualitative and quantitative indicators are included, consensus about how many and what kinds of indicators are valuable for fraudulent financial statement (FFS) detection can hardly reach. Indicators chosen varied with specific empirical data and application background. Vivid financial fraud cases nowadays manifest the limitations of this research paradigm. Financial statement fraud detection researches using quantitative variables and qualitative variables have already reached a plateau.

In contrast, few people have noticed the importance of the textual information for detecting fraudulent financial statement (also listed in Table 1). Cecchini et al. (2010b), Humpherys et al. (2011), and Larcker & Zakolyukina (2012) first extracted linguistic cues from textual information, and then applied data mining techniques to classify them into fraud and nonfraud cases. Glancy & Yadav (2011) developed a computational fraud detection model dealing with textual data in MD&A section.

We find that both the textual information and text mining techniques have not attached enough importance in detection of FFS. In addition, there are many problems pertinent to the existing text mining techniques as articulated in the Introduction section. Thus, we introduce a new text mining method that can not only settle the existing weaknesses but also more suitable for the detection of FFS.

3. RESEARCH METHOD AND THEORETICAL DEVELOPMENT

3.1 Statistical Language Model

In this paper, we introduce the statistical language model into realm of fraudulent financial statement detection. An SLM represents a probability distribution over a sequence of tokens that reflects how frequently it may occur (Jurafsky & Martin 2000). Tokens in a language model are comprised of both words and punctuation marks. We define the vocabulary set as all words in the corpus, denoted by ν. Consider a sequence *S*, $w_1 w_2 \cdots w_l$, where $l \geq 1$, each random variable in *S* takes a value in ν. Hence the joint probability of *S* can be modeled by (1).

$$P(S) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)\cdots P(w_l \mid w_1, \cdots, w_{l-1}) = \prod_{i=1}^{l} P(w_i \mid w_{i-N+1}, \cdots, w_{i-1}) \quad (1)$$

The premise of last equation in (1) is that it assumes the *i*th token to be dependent on only previous $(N-1)$ tokens. This is called a *N*-gram model (Brown et al. 1992). In practice, we always use an unigram, bigram, trigram model, corresponding to $N=1$, $N=2$, $N=3$.

The conditional probability in (1) can be estimated by Maximum Likelihood Estimation (MLE) in (2), i.e., dividing the occurrence frequency of a particular sequence by the occurrence frequency of a prefix.

$$p(w_i \mid w_{i-N+1}, \cdots, w_{i-1}) = c(w_{i-N+1}, \cdots, w_i) / c(w_{i-N+1}, \cdots, w_{i-1}) \quad (2)$$

Here $c(\cdot)$ denotes the occurrence frequency. Due to finite corpus, $c(w_{i-N+1}, \cdots, w_i)$ becomes very small, even zero, if *N* grows larger. That is why quadrigram model is seldom used. Under such circumstance, MLE could produce poor estimates, which is called a *sparse* problem. Fortunately, a useful technique, called *smoothing*, can reevaluate the zero-probability and low-probability *N*-grams, and assign them non-zero values. Among different smoothing techniques in literature (Chen & Goodman 1999; Chen & Rosenfeld 2000), we introduce a widely used one, i.e., Katz Back-off model. In order to show clearly, we use a trigram model as an illustration in (3).

$$\hat{p}(w_i \mid w_{i-2}, w_{i-1}) = \begin{cases} \dfrac{c(w_{i-2}, w_{i-1}, w_i) - d}{c(w_{i-2}, w_{i-1})} & \text{If } w_i \in \mathrm{A}(w_{i-2}, w_{i-1}) \\ \alpha(w_{i-2}, w_{i-1}) \times \dfrac{\hat{p}(w_i \mid w_{i-1})}{\sum_{w_i \in \mathrm{B}(w_{i-2}, w_{i-1})} \hat{p}(w_i \mid w_{i-1})} & \text{If } w_i \in \mathrm{B}(w_{i-2}, w_{i-1}) \end{cases} \quad (3)$$

Here $\hat{p}(\cdot)$ is the estimated probability. For any bigram (w_{i-2}, w_{i-1}), we define $\mathrm{A}(w_{i-2}, w_{i-1}) = \{w_i : c(w_{i-2}, w_{i-1}, w_i) > 0\}$, $\mathrm{B}(w_{i-2}, w_{i-1}) = \{w_i : c(w_{i-2}, w_{i-1}, w_i) = 0\}$. Back-off means if there is no particular trigrams in training corpus, its probability can be estimated by using the bigram probability instead. d is a constant discount value, such as 0.5. The discount method is to determine how much total probability mass to set aside for all the events we haven't seen. This total probability mass is estimated by (4).

$$\alpha(w_{i-2}, w_{i-1}) = 1 - \sum_{w_i \in \mathrm{A}(w_{i-2}, w_{i-1})} \frac{c(w_{i-2}, w_{i-1}, w_i) - d}{c(w_{i-2}, w_{i-1})} \quad (4)$$

Then this total probability mass is distributed by the back-off methods among the trigrams that not been observed. Perplexity is the most common metric used to evaluate *N*-gram model. The smaller the perplexity value, the better the model is. Suppose there are total *M* tokens, and total *m* sentences in the test corpus. The perplexity is then defined in (5).

$$\text{Perplexity} = 2^{-l}, \text{ where } l = (1/M) \sum_{i=1}^{m} \log p(s_i) \quad (5)$$

However, *N*-gram model is unable to capture large-span context, which limits its ability to estimate the occurrence probability of a target. For example, we cannot predict the token "fell" from token "stocks" in the sentence "stocks, as a result of the alleged financial fraud, fell sharply" only by a bigram or trigram model. In next section, we solve this problem by creating a model integrating framework of LSA with *N*-gram model discussed above.

3.2 Integrating Latent Semantic Analysis with N-gram Model

LSA is an automatic statistical approach to extract relations among words by mining their context of use in documents (Hofmann 2001; Landauer et al. 1998). Documents are viewed as semantically similar as long as they use similar words or the words tend to occur together.

Suppose vocabulary set ν contains *M* words, and training corpus includes *N* documents. A term-document matrix W ($M \times N$ matrix) is created, with rows corresponding to words in vocabulary set, and columns to documents. Each entry in the matrix is a weighted frequency of the corresponding

term in the corresponding document. However, the dimension of this matrix is usually very large, which causes heavy computation load. On the other hand, this matrix can be very sparse. Hence, the next step is to compress this matrix by singular value decomposition (SVD) in (6). Readers want to know details of how to construct this matrix and process of SVD can refer to Landauer et al. (1998).

$$W \approx \hat{W} = USV^T \quad (6)$$

Now the original $M \times N$ matrix W is decomposed into a reduced rank $M \times R$ term matrix U, a diagonal matrix of singular values S, and a $N \times R$ document matrix V. After SVD process, each row of V is a vector representation of the most important semantics of a particular document in R-dimensional space, so that noise information is filtered out. In addition, for a new document $\tilde{d}_p$, not in the training corpus, we can also compute its new document vector by $\tilde{v}_p = \tilde{d}_p^T US^{-1}$. We compare the semantic distance between any two documents by cosine of the angle between two corresponding rows.

Coccaro & Jurafsky (1998) proved LSA can be used as a language model. Let w_q denotes the word to be predicted and the context for this particular word is represented by H_{q-1}. In LSA, the context H_{q-1} always refers to the word history, i.e., current document up to word w_{q-1}, denoted by $\tilde{d}_{q-1}$. Thus, the LSA language model is given in (7).

$$p(w_q \mid H_{q-1}) = p(w_q \mid \tilde{d}_{q-1}) \quad (7)$$

Here $p(w_q \mid \tilde{d}_{q-1})$ is computed based on the cosine of angle between corresponding vector representation of w_q and $\tilde{d}_{q-1}$ in the R-dimension space. For details, please refer to Coccaro & Jurafsky (1998).

Since LSA and SLM focus on exactly different aspects in a document, e.g., LSA extracts the semantic similarity between documents, while SLM captures the words dependency relationships, an integration of them will create a great power to measure the similarity between documents (Bellegarda 2000; Coccaro & Jurafsky 1998). In the integrated model of SLM and LSA, the context H_{q-1} is comprised by both the previous N-gram (denoted by $H_{q-1}^{(n)}$) and history of this particular word ($H_{q-1}^{(l)}$). The integration model can be expressed as (8) theoretically.

$$\begin{aligned} p(w_q \mid H_{q-1}) &= p(w_q \mid H_{q-1}^{(n)}, H_{q-1}^{(l)}) = p(w_q \mid w_{q-N+1} \cdots w_{q-1}, \tilde{d}_{q-1}) \\ &= \frac{p(w_q \mid w_{q-N+1} \cdots w_{q-1}) \cdot [p(w_q \mid \tilde{d}_{q-1}) / p(w_q)]}{\sum_{w_i \in V} p(w_i \mid w_{i-N+1} \cdots w_{i-1}) \cdot [p(w_i \mid \tilde{d}_{i-1}) / p(w_i)]} \end{aligned} \quad (8)$$

Detailed computation of (8) can be found in Bellegarda (2000). In practice, the integrated model can be implemented by using LSA framework, with N-grams as terms, for simplicity. After SVD process, each document can be represented by a vector in low dimension semantic space, and then supervised or unsupervised learning methods can be utilized to classify documents into certain categories.

4. SAMPLE SELECTION AND DATA COLLECTION

In this paper, we try to detect the FFS cases in Chinese companies that went public in American capital market. We carry out the due diligence and check all Chinese companies listed in America one by one, including those have been halted, delisted, and private in the last decade. The detailed process about selecting companies with financial statement fraud and nonfraud companies is illustrated below.

4.1 Selecting Fraud Companies

We utilize Accounting and Auditing Enforcement Releases (AAERs), which are federal materials issued by SEC[2] for alleging accounting and/or auditing misconduct in U.S.-listed companies, to find Chinese companies involving fraudulent financial statements. First, we select all the AAERs related to Chinese companies. Second, a company reported in AAERs is viewed as a potential sample fraud company if the SEC accused top management of reporting misleading financial statements. Specifically, the SEC alleged violation of Rule 10(b)-5 of the 1934 Securities Exchange Act (Cecchini et al. 2010a; Fanning & Cogger 1998; Kaminski et al. 2004). Since SEC takes enforcement actions against companies, managers, auditors, and other parties involved in violations of SEC and federal rules, there are many charged objects. AAERs involving either public announcements or wrongdoing unrelated to financial misstatements, such as bribes or disclosure-related issues, and fraud in registration statements, are discarded in the third step (Dechow et al. 2011). We also ignore AAERs that report quarterly financial statement fraud mainly because it is unaudited (Brazel et al. 2009). In addition, we also exclude financial, insurance, and CPA companies because the accounting rules, asset valuations of these companies are different from other companies. Since a company could be alleged in multiple AAERs, we get the total fraud companies by discarding the duplicate AAERs. The amount of AAERs in each procedure is show in Table 2.

Total AAERs in the last decades (From 2003 to the first quarter of 2014)	1550
Total AAERs related to China-based companies	25
Less:	
AAERs not referencing violation of Rule 10(b)-5	0
AAERs involve either public announcements or wrongdoing unrelated to financial misstatements (such as bribes or disclosure-related issues, fraud in registration statements)	0
AAERs about quarterly financial statement fraud	0
Financial, insurance, and CPA companies	2
Subtotal companies	23
Less:	
Duplicate AAERs from the same company	6
Subtotal companies	17

Table 2. Detailed procedures for selecting AAERs about China-based companies

However, six of these selected companies are discarded from fraud samples due to lack of annual reports. Fich & Shivdasani (2007) suggested to create shareholder class action lawsuits to identify financial misstatements, which tend to be filed quickly after the disclosure of FFS. Thus, we employ class action lawsuits to supplement fraud sample in this paper. Fraud companied identified by class action lawsuits are filtered in SEC Litigation Releases (LRs), which follows almost the same procedures in searching AAERs. Note that lawsuit samples in LRs contain cases where fraud is alleged, but not actually proven. In order to make fraud samples identified by LRs more reliable, we choose the companies against which the lawsuits are eventually succeed. We add another six fraud samples from LRs, so the final data set includes a total of 17 fraud companies.

As fraud can span a number of years, we treated each year of fraud as fraud-year, ending up with a data set of 26 fraud-years. Detailed information about the fraud companies selected in this paper is shown in Table 3. Standard industry code (SIC) of the company indicates the industry it belongs to, and the total assets are collected in the first fraud year.

Num	Company name	Ticker	SIC	Total assets	Fraud year
1	China MediaExpress Holdings, Inc	CCME	7310	82,979,000	2009
2	China Natural Gas, Inc	CHNG	4923	247,448,555	2010
3	China North East Petroleum Holdings Ltd	NEP	1311	117730634	2009
4	Keyuan Petrochemicals, Inc	KEYP	2860	452,968,948	2010
5	Puda Coal, Inc	PUDA	1221	111,201,000	2009, 2010

[2] http://www.sec.gov/.

6	RINO International Corp	RINO	3569	118,627,800	2008, 2009
7	SUBAYE, Inc	SBAY	7374	48,104,000	2010
8	Universal Travel Group	UTRA	4700	49,426,680	2008~2010
9	China Sky One Medical, Inc	CSKI	2834	101,259,329	2007, 2008
10	ChinaCast Education Corp	CAST	8200	438,513,000	2010
11	AgFeed Industries Inc	FEED	0200	137,055,136	2008~2010
12	Yuhe International Inc	YUII	2015	76,763,036	2009, 2010
13	AutoChina International Ltd	AUTC	5500	500,206,000	2010, 2011
14	China Yuchai International Ltd	CYD	3510	819,532,000	2005
15	China Agritech, Inc	CAGC	2870	100,613,398	2009
16	Worldwide Energy & Manufacturing USA, Inc	WEMU	3990	34,650,123	2009
17	China Holdings, Inc	CHHL	2833	33,202,000	2008

Table 3. Detailed information about the selected fraud companies

4.2 Matching Nonfraud Companies

For each company in fraud sample, we match it with a control sample, a nonfraud company, for classification purpose. Nonfraud samples are matched with the fraud sample directly by using COMPUSTAT[3] on the basis of year, size, and industry.

First, the fiscal year is controlled to mitigate the effects of seasonal earnings patterns, concurrent economic conditions. We use the first fraud year that the company violates the lawsuits (e.g., prosecuted in AAERs or LRs) to make a comparison between fraud and nonfraud companies. Then, the firm size is measured by total assets. A nonfraud company is viewed as similar if total assets lie within the range of 30 percent of that for fraud company in the first fraud year (Kaminski et al. 2004). If no match is found, total sales are used. At last, the industry is controlled to mitigate industry characteristics. Nonfraud company is identified within the same four-digit SIC code as the fraud company. If no match is found, first three-digit code is used. Otherwise, first two-digit code is used.

After this matching process, 17 control companies (in other words, 26 firm-years), which are most similar and relevant in firm size and industry with the corresponding fraud companies in the first fraud year, are selected. Note that we cannot guarantee that companies in control group are non-fraudulent. We only guarantee that there is no publicly available financial fraud information about control samples so far. A control company should not be the one has submitted an amended financial statement, because amended financial statement is a signal of coming investigation of SEC (Dechow et al. 2011). In addition, companies that are covered by business press for fraudulent activities will be discarded.

5. DATA ANALYSIS AND EVALUATION

In this paper, we also used the Management's Discussion and Analysis (MD&A) section in annual report as analysis target. MD&A section is required to give investors a sense of management's perspective on the health and future outlook of a company. Prior researches have verified the ability of MD&A section for detecting financial statement fraud (Cecchini et al. 2010b; Glancy & Yadav 2011; Humpherys et al. 2011). In this paper, we only study the fraudulent statements in annual reports other than quarterly reports in that Form 10-Q is unaudited. Since some China-based companies are viewed as foreign private issuers by SEC, Form 20-F serves as annual report to provide information. Under this circumstance, we use a similar section, called Operating and Financial Review and Prospects (OFRP), in Form 20-F instead of MD&A section in Form 10-K or Form 10-KSB. In addition, all the Forms selected are the original Form other than the restated or amended ones.

[3] COMPUSTAT is a database of financial, statistical and market information on global companies through the world (www.compustat.com).

5.1 Data Preparation

A data preparation process is done in advance. First, we download the Form 10-K (or Form 10-KSB, Form 20-F) of each fraud company in fraud year(s) and the corresponding Form for each control company in the same year. We have 26 fraud sample and 26 nonfraud (control) sample in total. Second, the MD&A (or OFRP) section of each sample is extracted without the tables and figures, and is put into a text file individually. Third, all capital letters are changed to lowercase in each file. Fourth, each row in the file only contains one sentence. In addition, a vocabulary set is built using words in the training corpus for the construction of statistical language model, latent semantic analysis model, and integrated model. In the total 52 text files, there are 17,871 sentences, and 428,342 word tokens. The average number of word tokens in each sentence is 24.

When the data set is in position, we first build the statistical language model and latent semantic analysis model individually, and then build the integrated model, which includes four types of integration mode. Each model is illustrated in detail as follows.

5.2 Statistical Language Model

The SRI Language Modeling Toolkit (SRILM) (Stolcke 2002) is used to build the statistical language model for detecting financial statement fraud. Except for inserting a tag <s> in front of each sentence and a tag </s> in the end indicating the start and the end of a sentence, we leave all texts as they were to preserve original information. There are total 20,440 unigrams, 107,713 bigrams, and 69,917 trigrams in the data set.

Using a five-fold cross validation scheme, we build the statistical language model only using four-fifths samples each time, and use the rest samples for testing the performance of the model built. After five times cross validation, the average performance is used for comparison. In our experiment, there are 40 samples in training set, and 12 samples in testing set. Katz Back-off smoothing is used in each round of training. We use three classic evaluation metrics, i.e., accuracy, precision, and recall, to assess the testing performance. Accuracy is defined as the percentage of companies that are classified correctly. Precision is defined as the percentage of companies that are classified as fraud and are actually fraudulent. Recall is the percentage of fraudulent companies that are correctly classified.

In the training process, we build two statistical language models, one for fraud samples and the other for nonfraud samples in training set, which are called fraud model and nonfraud model respectively. In the evaluation step, the two trained models are used to compute the perplexity of each sample in testing set. The corresponding class label of the model with smaller perplexity value is the prediction result of this test sample. Thus evaluation metrics can be estimated by comparison between the true class label and the predicted label.

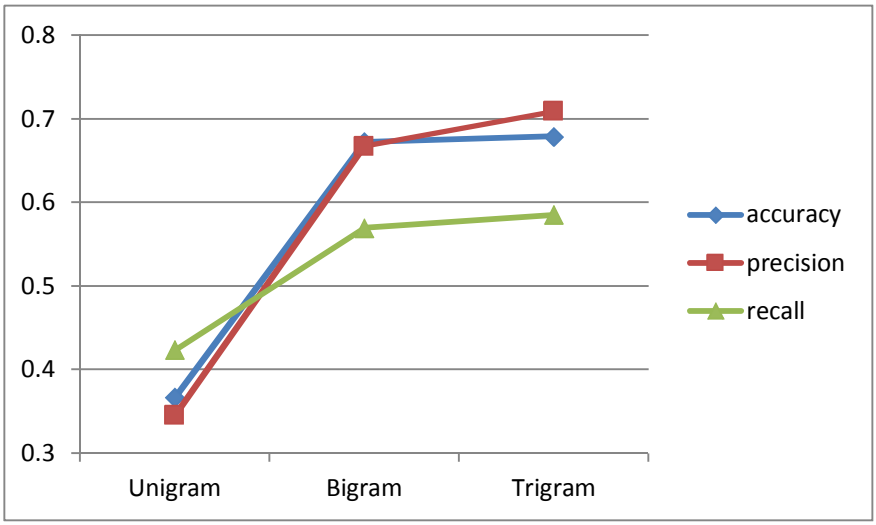

Figure 1. Results of Statistical language model

Given that quadrigram model is seldom used in practice, we test the unigram, bigram, and trigram model under this five-fold cross validation. The results are recorded in Figure 1. Under our data set, unigram model gets poor performance, even lower than 50 percent. Trigram model is a little better than bigram model, but the difference is hard to distinguish. Although bigram and trigram model are significantly superior to unigram model, the overall performance less than 70 percent is unsatisfying.

5.3 Latent Semantic Analysis Model

Latent semantic analysis model uses word tokens in vocabulary to build the term-document matrix. Sole LSA model is equal to the integrated model with only unigram as terms. The Genism Toolkit (Řehůřek & Sojka 2010) is used to build LSA model and integrated models. All punctuation marks are eliminated from the data set, and word tokens are stemmed using Porter stemmer.

In this paper, evaluations of LSA model and integrated models follow supervised learning paradigm. We use 40 samples in data set to train the model, and the rest 12 samples to test. We implement the training and testing process for 50 times, and use average accuracy to assess the performance. Four types of well-known supervised classification methods, i.e., support vector machine (SVM), neural network (NN), C4.5 decision tree (C4.5), and logistic regression (LR), are adopted in our experiment. Results of these classifiers for LSA model are shown in Figure 2.

As SVD is a dimension reduction process, each document can be represented in low dimension semantic space. We test different space dimensions to investigate whether the performance of the model is moderated by the dimension left. Since we only have 40 documents in training set, the maximum dimension is 40. The dimensions we choose in this paper range from 5 to 40. Final performances of LSA model with different combinations of dimension and classifier are shown in Figure 2.

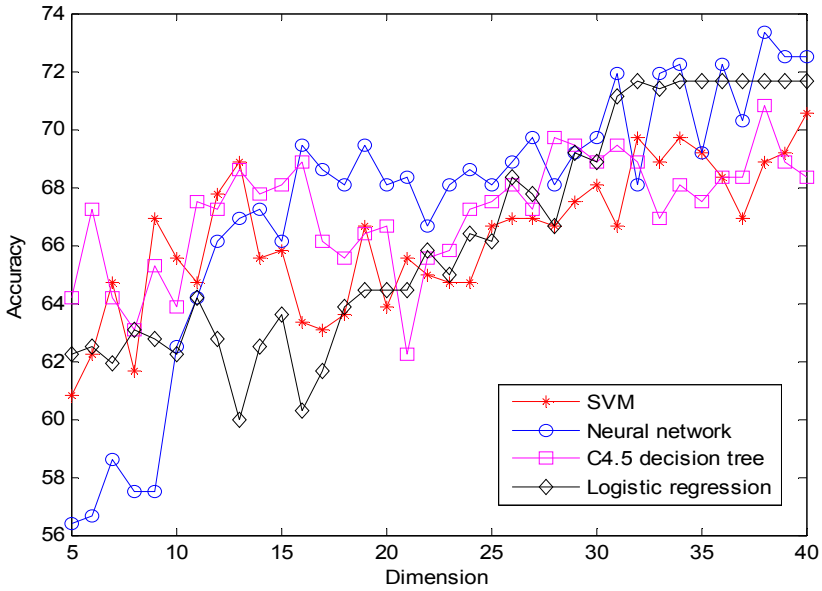

Figure 2. Results of latent semantic analysis model

The performance patterns of four classifiers are almost similar, with low performance in low dimension semantic space and relatively high performance in high dimension. This indicates the robustness of LSA model. Benefit from noise elimination in SVD process, NN, C4.5, and LR obtain the highest accuracy in dimension lower than 40. However, SVM achieves best accuracy at the highest dimension. Maybe it is because the dimension reduction loses too much information for SVM classifier. Performance of NN classifier fluctuates wildly with the increase of dimension, ranging from 56.38 percent to 73.33 percent. Comparing Figure 1 and Figure 2, we can clearly see LSA model significantly superior to unigram SLM. However, LSA has no comparable advantage to bigram model and trigram model, especially under low dimension.

5.4 Integrated Model

There are four types of integrated models: (1) LSA with only bigram as terms (we call it Model 1), (2) LSA with only trigram as terms (Model 2), (3) LSA with both unigram and bigram as terms (Model 3), (4) LSA with all unigram, bigram, and trigram as terms (Model 4). We follow the same analysis and evaluation process of LSA model in section 5.3. Results of first two models are shown in Figure 3, and results of last two models are reported in Figure 4.

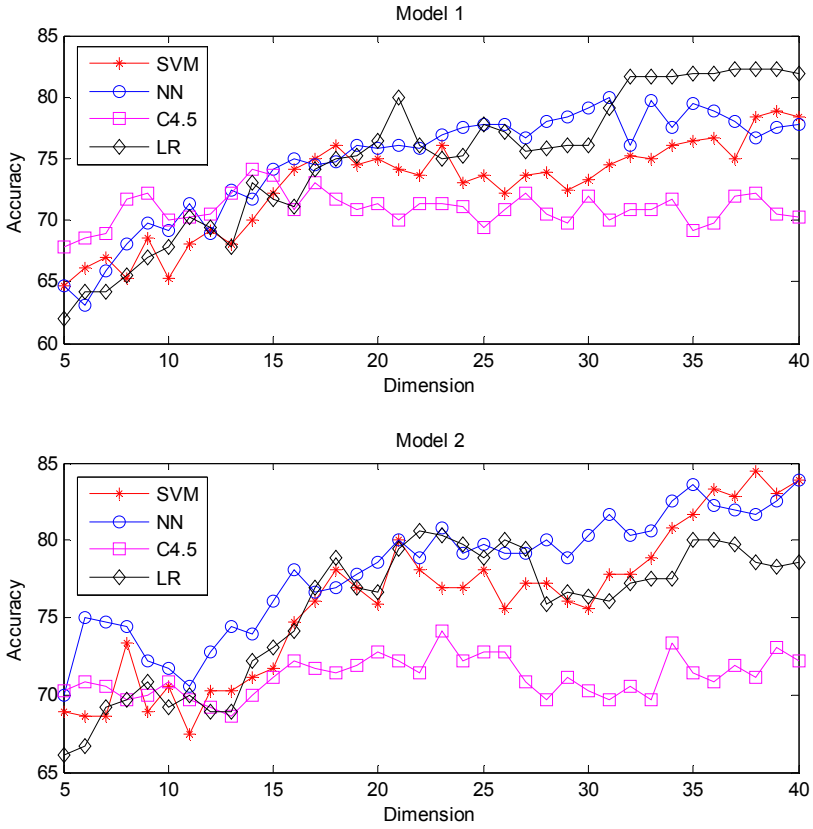

Figure 3. Results of Model 1 and Model 2

According to Figure 3 and Figure 4, four integrated models can achieve relatively high accuracy around 80 percent, among which Model 2 obtain the highest accuracy, i.e., 84.44 percent, using SVM classifier at 36 dimension semantic space. There is a growing trend of performance with the increase of space dimension. Some models achieve the best performance at maximum dimension while others not, which depends on the importance of information loss to classifiers. Similar trends exhibited in four classifiers verifies the robustness of the integrated model as well. Performances of four integrated models are much better than bigram and trigram SLM by comparing the accuracy can be achieved. In order to make a comparison of four integrated models together with sole LSA model, we list average accuracy of different dimensions for the five models under four classifiers in Table 4.

Model	SVM	NN	DT	LR
LSA	66.28	67.30	67.16	66.30
Model1	72.77	74.75	70.93	74.90
Model2	76.04	78.08	71.17	75.53
Model3	67.68	70.46	69.97	70.19
Model4	68.51	73.06	71.16	72.10

Table 4. Average accuracy of different models under four classifiers

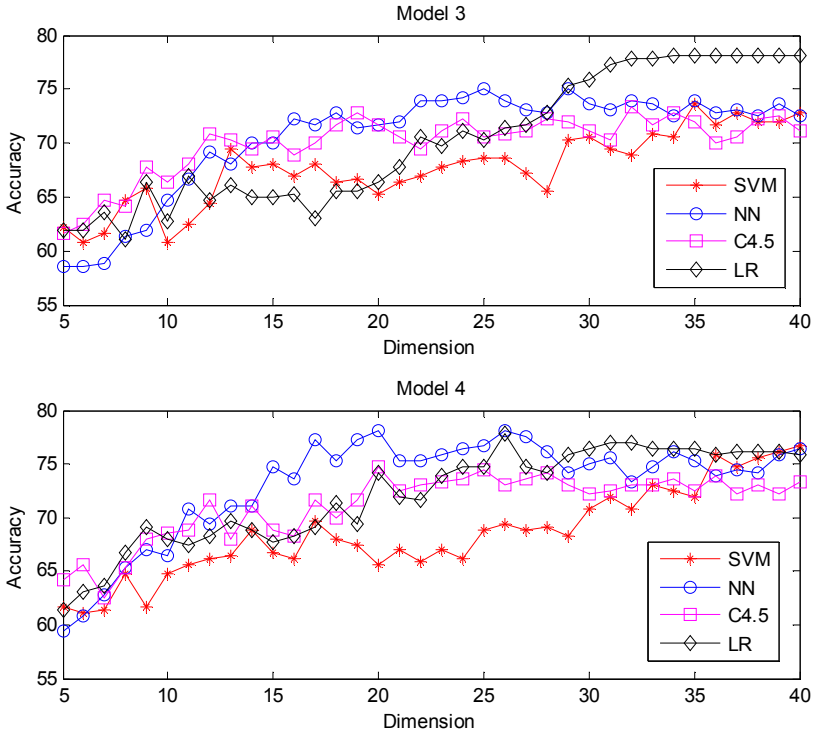

Figure 4. Results of Model 3 and Model 4

In Table 4, we clearly see that Model 2 obtains the highest accuracy for all classifiers used. In contrary, sole LSA model gets the lowest accuracy. In addition, taking SVM classifier as an example, we use a two-way ANOVA test to prove whether there is a significant difference in the performance of different models. In this test, five models in Table 4 and 36 dimensions act as independent variables, and accuracy as dependent variable. The result (F = 149.27, P value <0.001) demonstrates significant differences among models. Further Tukey post-hoc test is used to make multiple comparison between models. The result shows that Model 2 has significantly advantage over all the other models, and sole LSA model gets the worst performance. The performance of Model 1 is slightly worse than Model 2, followed by Model 3 and Model 4. There is no significant difference between Model 3 and Model 4. The findings are true for NN, C4.5, and LR classifiers as well. We can concludes that integrated model using trigram as terms is the best model under our circumstance, in other words, trigrams in financial statement contain relatively more information and useful patterns for classifying fraudulent and non-fraudulent ones. The high-performance model build in this paper to detect financial statement fraud is useful for stock market surveillance and public investors. Researchers and practitioners are suggested to pay more attention to investigation of textual information in financial statements, especially the trigram patterns.

6. CONCLUSIONS

In this paper, we introduce language model into realm of financial statement fraud detection. Four integrated models combining statistical language model and latent semantic analysis are built. Four supervised classifiers are used to test the performance and robustness of these models. In the empirical application to 52 China-based company firm-year samples, we demonstrate the superiority of integrated models than sole SLM and sole LSA model. In addition, via two-way ANOVA test, we find the integrated model using trigram as terms is the best one among four types of integrated models under our data set. The sensitive effect of dimension reduction on performance is also examined in this paper, showing a clear growing trend of performance with the increase of semantic dimension.

The integrated model has at least three outstanding properties compared to previous text mining researches in financial fraud detection. (1) The model uses all words in the text without relying on the

extraction of a predefined set of cues to fraud. (2) The model learns word dependencies relationships. (3) The model can help capture semantic relationships so as to approximate the meaning of sentences.

This research contributes to the field of financial fraud detection from both theoretical and empirical perspective. From the lens of theory, we build integrated language model to detect the strategic use of deceptive language in financial statements. In addition, by integrating LSA framework, we offset the shortcomings of SLM, i.e., the inability to capture long-span information. The integrated model can capture semantic patterns which distinguish fraudulent financial statements from non-fraudulent ones as well. From empirical perspective, we verify the effectiveness of integrated language model in detection of fraudulent financial fraud. The model developed in this paper can benefit market regulators and investors by successfully detecting financial statement fraud. Practitioners in stock market surveillance are suggested to pay more attention to textual information in financial statements, especially the bigram and trigrams patterns, for creating more effective regulation mechanisms.

Some researchers may argue the small data samples analyzed in this paper. However, we almost cover all the fraudulent firm-year samples from China-based companies in U.S. stock market. Our future research will expand the analysis process in this paper to all fraudulent companies listed in U.S.. We believe that will be a quite large data set.

Future research can also study the document structures, writing styles, genre, and sentiments in financial statement to create innovative mechanisms which are expected to have acceptable error rate and have better performance than existing methods.

ACKNOWLEDGMENTS

This research was supported by Research Grants Council of Hong Kong (Ref No. 193213).

References

Abbasi, A., Albrecht, C., Vance, A., and Hansen, J. (2012). Metafraud: a meta-learning framework for detecting financial fraud. Mis Quarterly, 36(4), 1293-1327.

Albrecht, S., and Romney, M. (1980). Auditing implications derived from a review of cases and articles relating to fraud. Proceedings of the 1980 Touche Ross. University of Kansas Symposium on Auditing Problems, 101-119.

Bankruptcydata.com (2012). 20 largest public domain company bankruptcy filings 1980 - present. (http://www.bankruptcydata.com/Research/Largest_Overall_All-Time.pdf, accessed April 8, 2013).

Bellegarda, J. R. (2000). Exploiting latent semantic information in statistical language modeling. Proceedings of the IEEE, 88(8), 1279-1296.

Beneish, M. D. (1999). The detection of earnings manipulation. Financial Analysts Journal, 24-36.

Brazel, J. F., Jones, K. L., and Zimbelman, M. F. (2009). Using nonfinancial measures to assess fraud risk. Journal of Accounting Research, 47(5), 1135-1166.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. Computational linguistics, 18(4), 467-479.

Cecchini, M., Aytug, H., Koehler, G. J., and Pathak, P. (2010a). Detecting management fraud in public companies. Management Science, 56(7), 1146-1160.

Cecchini, M., Aytug, H., Koehler, G. J., and Pathak, P. (2010b). Making words work: using financial text as a predictor of financial events. Decision Support Systems, 50(1), 164-175.

Chen, S. F., and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. Computer Speech & Language, 13(4), 359-393.

Chen, S. F., and Rosenfeld, R. (2000). A survey of smoothing techniques for ME models. Speech and Audio Processing, IEEE Transactions on, 8(1), 37-50.

Coccaro, N., and Jurafsky, D. (1998). Towards better integration of semantic predictors in statistical language modeling. ICSLP, Citeseer.
Coderre, D. G. (1999). Fraud detection: using data analysis techniques to detect fraud Global Audit Publications.
Dechow, P. M., Ge, W., Larson, C. R., and Sloan, R. G. (2011). Predicting material accounting misstatements. Contemporary Accounting Research, 28(1), 17-82.
Deshmukh, A., Romine, J., and Siegel, P. H. (1997). Measurement and combination of red flags to assess the risk of management fraud: a fuzzy set approach. Managerial Finance, 23(6), 35-48.
Deshmukh, A., and Talluru, L. (1998). A rule-based fuzzy reasoning system for assessing the risk of management fraud. International Journal of Intelligent Systems in Accounting, Finance & Management, 7(4), 223-241.
Fanning, K. M., and Cogger, K. O. (1998). Neural network detection of management fraud using published financial data. International Journal of Intelligent Systems in Accounting, Finance & Management, 7(1), 21-41.
Fich, E. M., and Shivdasani, A. (2007). Financial fraud, director reputation, and shareholder wealth. Journal of Financial Economics, 86(2), 306-336.
Gaganis, C. (2009). Classification techniques for the identification of falsified financial statements: a comparative analysis. Intelligent Systems in Accounting, Finance & Management, 16(3), 207-229.
Glancy, F. H., and Yadav, S. B. (2011). A computational model for financial reporting fraud detection. Decision Support Systems, 50(3), 595-601.
Green, B. P., and Choi, J. H. (1997). Assessing the risk of management fraud through neural network technology. Auditing: A Journal Of Practice & Theory, 16(1), 14-28.
Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. Machine learning, 42(1-2), 177-196.
Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., and Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. Decision Support Systems, 50(3), 585-594.
Jurafsky, D., and Martin, J. H. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition Pearson Education India.
Kaminski, K. A., Wetzel, T. S., and Guan, L. (2004). Can financial ratios detect fraudulent financial reporting? Managerial Auditing Journal, 19(1), 15-28.
Kirkos, E., Spathis, C., and Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications, 32(4), 995-1003.
Kotsiantis, S., Koumanakos, E., Tzelepis, D., and Tampakas, V. (2006). Forecasting fraudulent financial statements using data mining. International Journal of Computational Intelligence, 3(2), 104-110.
Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284.
Larcker, D. F., and Zakolyukina, A. A. (2012). Detecting deceptive discussions in conference calls. Journal of Accounting Research, 50(2), 495-540.
Lin, J. W., Hwang, M. I., and Becker, J. D. (2003). A fuzzy neural network for assessing the risk of fraudulent financial reporting. Managerial Auditing Journal, 18(8), 657-665.
Liou, F.-M. (2008). Fraudulent financial reporting detection and business failure prediction models: a comparison. Managerial Auditing Journal, 23(7), 650-662.
Loebbecke, J., and Willingham, J. (1988). Review of SEC accounting and auditing enforcement releases. Unpublished working paper. University of Utah.
Pasiouras, F., Gaganis, C., and Zopounidis, C. (2007). Multicriteria decision support methodologies for auditing decisions: the case of qualified audit reports in the UK. European Journal of Operational Research, 180(3), 1317-1330.
Perols, J. (2011). Financial statement fraud detection: an analysis of statistical and machine learning algorithms. Auditing, 30(2), 19-50.
Persons, O. S. (2011). Using financial statement data to identify factors associated with fraudulent financial reporting. Journal of Applied Business Research, 11(3), 38-46.

Ravisankar, P., Ravi, V., Raghava Rao, G., and Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. Decision Support Systems, 50(2), 491-500.
Řehůřek, R., and Sojka, P. (2010). Software framework for topic modelling with large corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 45-50.
Spathis, C., Doumpos, M., and Zopounidis, C. (2002). Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariate statistical techniques. European Accounting Review, 11(3), 509-535.
Spathis, C. T. (2002). Detecting false financial statements using published data: some evidence from Greece. Managerial Auditing Journal, 17(4), 179-191.
Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. INTERSPEECH.
Summers, S. L., and Sweeney, J. T. (1998). Fraudulently misstated financial statements and insider trading: an empirical analysis. Accounting Review, 131-146.
Wang, J.-H., Liao, Y.-L., Tsai, T.-m., and Hung, G. (2006). Technology-based financial frauds in Taiwan: issues and approaches. SMC, 1120-1124.
Yuan, J., Yuan, C., and Deng, X. (2008). The effects of manager compensation and market competition on financial fraud in public companies: an empirical study in China. International Journal of Management, 25(2), 322-335.
Zhou, L., Shi, Y., and Zhang, D. (2008). A statistical language modeling approach to online deception detection. Knowledge and Data Engineering, IEEE Transactions on, 20(8), 1077-1081.