COMBINING FINANCIAL RATIO AND LINGUISTIC ANALYSES

TO DETECT FRAUDULENT FINANCIAL STATEMENTS

By

Daniel Ferguson

RECOMMENDED:

_____
Yonggang Lu, Ph.D.

_____
Clare Dannenberg, Ph.D.

_____
Gökhan Karahan, Ph.D.
Chair, Advisory Committee

_____
Bogdan Hoanca, Ph.D.
Director of Graduate Programs

APPROVED:

_____
Rashmi Prasad, Ph.D.
Dean, College of Business & Public Policy

_____
Helena Wisniewski, Ph.D.
Vice Provost for Research and Graduate Studies
Dean of the Graduate School

_____
Date

COMBINING FINANCIAL RATIO AND LINGUISTIC ANALYSES TO DETECT FRAUDULENT FINANCIAL STATEMENTS

A

THESIS

Presented to the Faculty

of the University of Alaska Anchorage

in Partial Fulfillment of the Requirements

for the Degree of

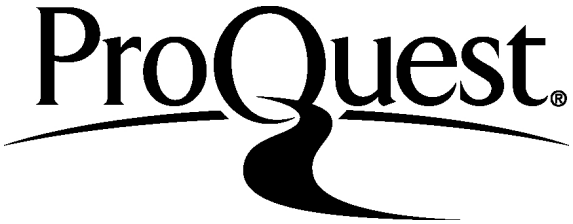
MASTER OF BUSINESS ADMINISTRATION

By

Daniel Ferguson, B.A.

Anchorage, Alaska

August 2016



Abstract

Fraud continues to be an increasingly important topic in business, particularly during financial statement audits. While still a new concept, data-mining techniques have become increasingly popular for determining where to allocate resources throughout an audit. This study investigated whether combining financial ratio and linguistic analyses provides better predictive results than using either analysis alone. The hypothesis was tested utilizing logistic regression as well as artificial neural networks and random forest analyses with a sample of 110 annual financial reports. Results showed that the Combined Model performed better than both the Financial and Linguistic models in four out of six tests. The combined model also had a lower Akaike Information Criterion and Bayesian Information Criterion when compared to the other two models for all tests. It also appeared that the linguistic variables Reward, Risk and Power had significant predictive ability, a relatively novel idea that can be explored in future studies.

Table of Contents

List of Figures

Page

List of Tables

Page

Acknowledgments

I would like to thank my Thesis Advisor Committee, Dr. Gökhan Karahan, Dr. Yonggang Lu, and Dr. Clare Dannenberg for their continued assistance and guidance throughout this project.

I would also like to thank Dr. Bogdan Hoanca, Dr. Rashmi Prasad, and Dr. Helena Wisniewski for reviewing the manuscript and providing additional comments, as well as Elisa Mattison for providing formatting assistance.

I would like to thank the UAA College of Business and Public Policy, English Department, and Graduate School for their assistance and support, as well as the Society of Business, Industry and Economics for accepting the draft of this paper for presentation at their annual 2016 SOBIE Conference.

Finally, I would like to thank my family, friends and especially Işıl Aşkın for their support and assistance through many long hours of research and writing during the project.

Introduction

Fraud has received increased scrutiny in recent years since the passing of the Sarbanes Oxley Act and numerous high-profile cases such as Enron and WorldCom (Benston, 2003). Fraud is found in many different aspects of the business environment; while the larger cost of fraud comes from large corporations, most fraud cases exist in smaller companies with less than 100 employees.

It has been previously estimated that fraud costs companies approximately $400 billion dollars a year in the U.S. alone (Wells, 2007), with more recent estimates increasing that number to $572 billion dollars per year (Association of Certified Fraud Examiners, 2008). In a recent study, PKF Accountants LLP predicted global fraud according to the Improper Payments Information Act (IPIA) standards[1] and found that the yearly average global cost of fraud as a percentage of income for 2012-2013 was estimated at approximately $4.23 trillion per year (PKF Littlejohn LLP, 2015). The same study also notes both the cost and prevalence of fraud appears to be increasing. Furthermore, the Association of Certified Fraud Examiners found that financial statement fraud is the most expensive with a median cost of $975,000 per case, over four times higher than the second most expensive type of fraud (Association of Certified Fraud Examiners, Inc, 2016). This concerning trend indicates the importance of finding new ways to detect fraudulent financial reporting during or before the initial audit rather than years later.

There are a number of sources that offer their own definition of fraud. The Oxford English Dictionary defines fraud as "wrongful or criminal deception intended to result in financial or personal gain" (Pearsall, 1999), while the FBI defines fraud as "comprising of deceit,

[1] IPIA standards require estimates to utilize a sample size large enough to give a 90% confidence interval +/- 2.5%. Many countries in Europe however are required to utilize a sample size large enough to give a 95% confidence interval +/- 1%.

concealment, and/or violation of trust" (Dutta, 2013). For this particular study, I focus on the definition of financial statement fraud given by the American Institute of Certified Public Accounts as "The intentional misstatements or omissions of amounts or disclosures in financial statements" (American Institute of Certified Public Accountants (AICPA), 1997) as well as the idea of management fraud, defined by Elliot and Willingham (1980) as "Deliberate fraud committed by management that injures investors and creditors through misleading financial statements." It is important to note that management fraud is generally considered more difficult to detect when compare to other types of fraud because in many cases management directly communicates with auditors about the company which gives those perpetrating fraud the ability to intentionally and directly mislead auditors (Beasley, 1996).

When looking at the changes in responsibility of auditors in detecting fraud, particularly after SAS No. 82 and superseding SAS No. 113, the responsibility of auditors has been both increased and clarified. An important change is that auditors must provide "reasonable assurance about whether the financial statements are free from material misstatements" (AICPA c, AU 316.02). Even though later SAS publications have clarified that auditors are unable to provide absolute assurance that financial statements are free from material misstatements, auditors must be able to provide reasonable assurance and proof of due diligence, including requiring the use of analytical procedures during an audit (American Institute of Certified Public Accountants (AICPA), 1988).

While still a relatively new concept, the use of data-mining techniques has become popular for a number of financial applications. In particular, data-mining has been successfully used as a tool to assist auditors during all stages of the audit, particularly during the planning stage when determining where to allocate often limited resources (Fraser, Hatherly, & Lin, 1997).

Phua, Lee, Smith & Gayler (2005) state that fraud detection has become one of the most successful uses of data-mining techniques in both the industry and government environment. Historically, data mining has focused on using logistic regression and quantitative data such as financial statement ratios to detect fraud (Kaminski, Wetzel, & Guan, 2004). Recently however, data mining techniques have expanded to include numerous advanced analytical procedures such as artificial neural networks (ANN) and qualitative data such as the composition of corporate board members (Beasley, 1996; Kirkos, Spathis, & Manolopoulos, 2007; Uzun, Szewczyk, & Varma, 2004).

Current Use of Data Analytics

While the field of data analytics and in particular the idea of "big data" are both relatively new, the use of data analytics by auditing companies has been steadily increasing for some time. Beginning in the early 1980's, the AICPA released a statement noting that "Analytical review procedures may be performed in the initial planning stages, during the examination, and at or near the conclusion of an audit" (AICPA, 1983, AU 318). Following AU 318, SAS No. 56 required the use of analytical procedures in the planning and reviewing stages of audits "consisting of evaluations of financial information made by a study of plausible relationships among both financial and non-financial data" (AICPA 1988b).

Historical research on the use of data analytics in fraud examinations began not long after the release of AU318. Tabar and Willis (1985) examined the literature and found an observable increase in the use of more advanced data analytic methods throughout the years immediately following AU318. They also interviewed seven audit managers from one of the "Big Eight" audit firms, with each manager selecting two audit clients to use in the data analysis. Results indicated that there was a significant increase in the amount of advanced data analytics

procedures used by the auditing firm between 1978 and 1982, with a corresponding significant decrease in the use of non-quantitative procedures, particularly in the planning stages of an audit. All seven auditors stated that they believe advanced analytics procedures would increase throughout the future of auditing.

While research over the past few decades has illustrated an increase of interest in the use of data analytics for auditing, there are relatively few case-studies of data analytics in current literature. An article published by Bump (2015) outlines the current strategy for the State of Massachusetts Auditor's Office to use data analytics as one of their core processes. Beginning in 2012, the plan has consisted of three phases: (1) Build a test environment (2) Consolidate gains and analytical algorithms (3) Finalize analytics engine and train staff. Currently in the second phase of the project, the auditing office has already seen improvement in their workflow while using data analytics. Bump states that the analytics software has identified over $20 million in questionable spending, including 1,164 Social Security payments totaling $2.39 million which were paid to "deceased" individuals 6-27 months after their death.

Current Literature on Data Analytics

Throughout the literature there seems to be a number of major models types used in fraud detection. The current study focuses on four: logistic regression, artificial neural networks, decision tree analysis, and random forest analysis.

Analytical models.

Logistic regression. Logistic regression for auditing takes the form of a binary dependent variable (fraud or no-fraud) that is predicted by multiple independent variables. Logistic regression appears to be one of the first methods tested for detecting fraudulent financial statements (FFS). Zopounidis and Doumpos (1999) used logistic regression to detect falsified

financial statements utilizing only financial ratios with a success rate of 84%. In a related study, Spathis, Zopounidis and Doumpos (2002) used a multi-criterion decision aid method (MCDA) called UTADIS (UTilities Additives DIScriminants) to predict FFS with a sample of 76 Greek firms. Input variables for the model included Sales / Total Assets, Net Profit / Sales, Inventory / Sales, and Total Debt / Total Assets. They found that the UTADIS method performed significantly better at detecting FFS than their selected benchmarks of Discriminant Analysis and Logit Analysis methods.

Artificial neural networks. Inspired by biological neural networks, artificial neural networks estimate the probability of an event occurring (in this case a binary fraud or no-fraud event) based on a large number of inputs. Neural networks tend to have a major advantage over other statistical methods for a number of reasons: They are adaptive, able to generalize to other data sets and do not require rigid assumptions such as normality of data. Neural networks also appear capable of reducing false negative error rates without a subsequent increase in false positive error rates, a relatively major drawback found in many other statistical methods (Green & Choi, 1997). Neural networks have already been successfully used in a number of other financial situations, such as bankruptcy prediction and assisting with credit approval decisions.

Artificial neural networks are not without their drawbacks; one major problem is they cannot be quantified like other statistical measures as there is no way to determine the significance level of a neural network's model at predicting FFS (Green & Choi, 1997). A neural network's architecture is also primarily the result of trial and error, meaning there is no quantifiable way to determine if the final architecture is an optimal solution. Since the AICPA requires the use of data analytics that have a proven logical basis, an auditor who used artificial neural networks as the primary analytical technique during an audit may be hard-pressed to

provide this proof if financial statements were found to be fraudulent after the auditor gave an unqualified opinion.

There are a number of studies that have used artificial neural networks to detect FFS. Green and Choi (1997) were able to construct three neural networks utilizing only five financial ratios and three financial accounts. Their data sample consisted of 113 FFS matched with 113 non-FFS that were filed with the SEC and selected directly from COMPUSTAT. A simple prior-year percentage change was used for the financial ratios and accounts as inputs for the model. Results indicated that all three models had a summed error rate of less than 1, of which the PSYDYNN model had the lowest error rate of 37.04%. Kirkos, Spathis and Manolopoulos (2007) found that while an artificial neural network was able to correctly classify 100% of fraud and non-fraud cases in a training data set, the model was able to correctly classify only 82.5% of fraud and 77.5% of non-fraud cases in a validation set. In this particular study, the artificial neural network was outperformed by a Bayesian belief model utilizing the same data inputs. This is an example of the tendency for artificial neural networks to "over-fit" a training dataset which could jeopardize the model's ability to generalize appropriately with novel data sets.

Decision tree analysis. Another statistical method that has been prevalent in the literature, decision tree analysis consists of a flowchart with a series of "branches" and "nodes." Each node consists of a probability test which, based on the outcome, determines which branch the model flows to. By utilizing a large number of branches and nodes, the model is able to predict the overall probability of a binary event based on a number of inputs.

In one study, Chen, Yeong-Jia, Goo & Shen (2014) created three different models utilizing logistic regression, stepwise regression, support vector machine and decision tree analysis to detect FFS. Out of the four models, the decision tree analysis reported the greatest

accuracy in detecting FFS (Chen et al., 2014). Results from another study reported that a random forest decision tree model was able to detect FFS with an accuracy of 88% when using eight input variables (Liu et al., 2015). In yet another study, Kirkos, Spathis and Manolopoulos were able to create a decision tree analysis that correctly classified 75% of fraud cases and 72.5% of non-fraud cases from a training sample of 72 cases (2007).

Random forest analysis. The final statistical method used in this study, random forest analysis is very similar to decision tree analysis except that it grows a very large number of trees without any pruning and then uses the aggregate score of each tree as a "vote" in determining the answer to a dichotomous variable. First proposed by Breiman (2001) and further developed by Liaw and Weiner (2002), random forest analyses have been used in a number of classification studies with good results.

The primary advantage to random forest analysis as opposed to a using a single decision tree is that as the number of decision trees utilized increases, the false negative, false positive, and out-of-bag error (the ability to predict novel information) rates tend to decrease and finally stabilize to a specific percentage. While a random forest can be anywhere from 100 trees to thousands, the average random forest is less than 10,000 decision trees. When looking specifically at fraud, Chengwei, Yixiang, Syed, and Hao (2015) created a random forest analysis utilizing seven variables and 500 decision trees that was able to correctly identify 88.41% fraudulent cases and 87.50% of non-fraudulent cases from a sample of 298 manufacturing companies on the Chinese Stock Exchange.

Linguistic analysis models.

Linguistic analysis software. There are a number of computer programs that have been used in studies specifically to detect falsified statements based solely on linguistic cues. Many

studies rely on automated computer packages such as the Agent99Analyzer, a program specifically created to detect falsified statements from linguistic cues in text and video (Fuller, Biros, Twitchell, Burgoon, & Adkins, 2006). In particular, Agent99Analyzer uses a module called "GATE" (General Architecture for Text Engineering) as a sub-tool for processing language. Three other popular software packages utilized in the literature include iSkim and CueCal (Zhou L. , Burgoon, Nunamaker, & Twitchell, 2004; Zhou L. , Burgoon, Twitchell, Qin, & Jr, 2004), as well as the Linguistic Inquiry and Word Count software (Hancock, Curry, Goorha, & Woodworth, 2008).

Linguistic credibility analysis for lie detection. There appears to be a sizeable body of literature on the use of computer-assisted analysis to detect lying in numerous types of communication, including text. One study analyzing 242 transcripts found that deceptive statements appear to use significantly more words in general, more sense-based words, more other-oriented words and finally fewer self-oriented words (Hancock, Curry, Goorha, & Woodworth, 2008). Another study utilized a meta-analysis to provide support that when compared to truthful statements, liars appear to experience more cognitive load, express more negative emotions, distance themselves more from events, expressed fewer sensory-perceptual words and referred less often to cognitive processes (Hauch, Blandon-Gitlin, Masip, & Sporer, 2015).

Software analysis has been used to analyze more than just text-based communications. In one study, Newman, Pennebaker, Berry and Richards (2003) analyzed five different contextual groups: Video abortion, typed abortion, written abortion, video with friends and video with mock crime. Transcribers translated written and verbal work into 568 samples and utilized the Linguistic Inquiry and Word Count (LIWC) software for text analysis. The study was able to

correctly classify false statements from true statements at a rate of 67% when the topic was constant and 61% when the topic was varied. A logistic regression equation was developed with five predictor variables that were used across all samples: First-person singular pronouns, third-person pronouns, negative emotion words, exclusive words and motion verbs.

Linguistic credibility analysis for financial statements. While there are a number of articles focusing on the automated detection of falsified statements in general text, there are relatively few articles that use linguistic analysis particularly for the detection of FFS.

Humpherys et al. (2010) hypothesized that FFS would contain a higher number of linguistic cues than non-FFS which could be detected using the Agent99Analyzer, a part-of-speech text analysis program. The 24 variables used in their study were all based on eight linguistic cues found to be significantly correlated with deceptive language during research by Zhou, Burgoon, Nunamaker & Twitchell (2004): Affect, complexity, diversity, expressivity, non-immediacy, quantity, specificity and uncertainty. After analyzing a sample of 202 10-K reports that were filed with the SEC, Humpherys et al. was able to use a Naïve Bayes classifier and a C4.5 decision tree classifier on a reduced 1ten variable model to detect FFS with an accuracy of 67.3%.

Use of Quantitative and Qualitative Data in Financial Statements

While I could find no specific examples of combining both linguistic credibility and financial ratio analyses to detect FFS, there are some examples of using both analyses to predict financial performance. One study analyzed both financial ratios (quantitative) and textual data (qualitative) in quarterly reports published by Nokia, Motorola and Ericsson from 1995 to 1999. Financial variables included profitability ratios, liquidity ratios, solidity ratios and efficiency ratios and were utilized in tandem with textual data from quarterly reports to predict financial

performance using SOM_PAK self-organizing maps. The study found that there is a time-lag between the two analyses, with the linguistic analysis appearing to forecast financial performance for the next quarterly report with a limited degree of accuracy (Kloptchenko, Eklund, Karlsson, Vanharanta, & Visa, 2004).

A follow-up study utilizing seven financial ratios and linguistic data from quarterly reports spanning 2000-2001 found similar results with the linguistic analysis providing a forecast of financial performance in later reports (Magnusson, et al., 2005). These two studies provide evidence that there is a connection between linguistic and financial data in quarterly reports that could reveal additional information if analyzed appropriately.

Role of the Current Study

In the previous literature review there have been numerous studies conducting only quantitative data analyses in order to detect falsified or FFS with limited success. While there appears to be a healthy body of literature on the automatic detection of falsified statements in text, I found only one study thus far using linguistic analysis for detecting falsified financials statements. There also appears to have been no studies to date attempting to combine both quantitative and qualitative methods specifically for fraud detection, presenting a gap in the literature.

Of particular interest is the difference between the type of errors that can occur when attempting to determine if a financial statement is fraudulent or not. The literature classifies a false positive as predicting a statement does contain fraud when it is actually does not, and a false negative as predicting a statement does not contain fraud when in fact it does. As can be expected, it is far costlier for an auditing firm to predict that a financial statement does not contain fraud when actually it does. When determining the accuracy of a predictive model,

accuracy in regards to false negative errors is generally considered the most important aspect for the model, whereas false positive errors are generally considered less important.

The current study attempts to fill a gap in current research by determining if combining both quantitative (financial ratios) and qualitative (linguistic cues) data analyses performed on financial statements filed with the SEC will significantly improve the ability to detect fraud when compared to using either quantitative or qualitative analysis in isolation. In order to test for this, I proposed the following hypotheses:

Hypothesis 1:

H_a: An analytical model that combines both financial (quantitative) and linguistic (qualitative) analyses will predict fraudulent cases better than an analytical model using only one of these data types.

H_o: There will be no perceptible difference in predictive power when comparing analytics utilizing the Combined Model with either the Financial or Linguistic Models.

Hypothesis 2:

H_a: Due to the nature of linguistic data being able to detect the intent to deceive, I also predict that the combined model will perform better than the financial model on detecting false negative errors.

H_o: There will be no perceptible difference in predictive power for false negative errors when comparing analytics utilizing the Combined Model with the Financial Model.

Methodology

Data Collection

The sample included a total of 110 publicly traded firms registered with the SEC, 55 that submitted fraudulent financial statements (FFS) and 55 that submitted non-fraudulent financial statements (N-FFS). Public auditors issued an opinion on the financial statements for each company with the exclusion of one that had forged an auditing report.

I considered a financial statement to be fraudulent if it had successful litigation from the SEC noting that (1) fraud had occurred and (2) the financial statements in question were materially misstated. After collecting all completed SEC litigations with a filing date beginning January 1, 2006 to December 31, 2015, the study began with a sample of 191 FFS. For the purpose of this study, I included both 10-K and 10-KSB documents since the differences between the two are considered to have a negligible effect on the results. I then narrowed this sample according to a number of qualifiers:

1. The company had common stock publicly traded on a major stock exchange.
2. The company's financial statements included both inventory and receivable accounts to allow for a standardized financial ratio analysis applicable to the entire sample.
3. Litigation from the SEC was completed and not pending any further investigation.

After applying these conditions, the FFS sample was narrowed down to a total of 55 observations to be used in the current study. Each FFS was then matched with a company that satisfied the following conditions:

1. Filed a 10-K or 10-KSB document with the SEC during the same year(s) as the matching company with FFS.
2. Did not have any prior or future fraud cases pending or completed with the SEC.

3. Originated from the same Standard Industrial Classification (SIC) code as the company with FFS.

I attempted to match fraudulent companies with non-fraudulent companies that had a difference in total assets of less than 20%, however this was not always possible due to a lack of companies to select from in certain SIC classifications. As can be seen in Table 1, the sample included an aggregate value with a difference of 22.56% in total assets between fraud and non-fraud companies. Standard deviations also appeared quite high when compared to their corresponding average, demonstrating a large amount of variability in both samples.

Table 1

Descriptive Statistics for FFS and N-FFS Samples

	FFS		N-FFS	
Measure	***M***	***SD***	***M***	***SD***
Total Assets	$ 1,098,966,351.71	$ 1,419,194,706.73	$ 1,450,848,956.10	$ 3,264,829,252.20
Total Liabilities	$ 673,391,249.67	$ 895,432,630.35	$ 994,105,327.11	$ 2,339,650,965.72
Total Sales	$ 1,017,585,657.78	$ 1,538,916,848.00	$ 1,459,051,001.00	$ 4,238,205,579.28
EBIT	$ 5,742,494.51	$ 236,615,987.45	$ 102,577,279.93	$ 637,957,817.79
Net Income	$ 3,034,551.87	$ 157,520,424.80	$ 108,504,981.29	$ 460,017,503.75
Market Cap	$ 1,014,617,573.13	$ 2,191,148,577.80	$ 1,942,661,648.80	$ 5,522,227,576.06
CS Outstanding	101,317,599.84	80,146,764.82	19,5524,371.73	192,143,456.22

Note: Common Stock Outstanding and Market Capitalization were obtained from reported amounts in the company's filing document.

Variable Selection

Financial ratio variables. Financial variables used in this study were selected based on a number of articles found in prior literature referenced in the introduction section. Unfortunately, previous literature had not agreed on any unified set of variables that should be used to detect FFS; many variables significant in one study were insignificant in another. Due to this, it became difficult to select a specific set of variables without including many found to be significant at least once in previous literature.

Prior literature and expectations. According to prior studies, I expected variables capable of measuring financial distress to be more predictive of FFS when compared to other financial ratio variables used (Phua, Lee, Smith, & Gayler, 2005). While indicators of higher sales and profits tend to signify a healthy company, it appears from prior literature that this was not always the case, possibly due to many fraudulent companies inflating revenues and causing higher sales or profit margins in order to make a company in distress appear otherwise healthy (Ravinskar, Ravi, Rao, & Bose, 2011). Therefore, while higher levels of profit or sales may indicate a lower chance of FFS, I did not expect variables that measure sales or profits to perform as well as measurements of financial distress such as debt or solvency ratios. Finally, the well-known Altman's Z-Score had also been a significant predictor in prior studies; I expected data from the current study to replicate these results (Spathis, Doumpos, & Zopounidis, 2002).

With the above expectations in mind, I approached the financial variable pool by including 25 financial ratios that were found to be significant in previous literature. I categorized the variables into seven specific categories that appeared to predict FFS: Solvency, growth, financial distress, liquidity, receivables, inventory, profitability, and some additional "red flags" that were found to be predictive in previous literature. These variables are summarized in Table 2 and include descriptive statistics for fraudulent and non-fraudulent samples.

Table 2

Financial Variables with Descriptive Statistics

Dimension	FFS		N-FFS	
I. Solvency	***M***	***SD***	***M***	***SD***
Long Term Debt / Total Assets	0.21	0.19	0.20	0.24
Total Debt / Total Assets	0.61	0.42	0.90	0.29
Debt / Equity	1.10	1.01	1.89	1.82
Logarithm of Total Debt	8.09	7.83	0.97	0.95
II. Growth				
Current Sales – Prior Year's Sales / Prior Period Sales	0.44	0.13	0.58	0.24
III. Financial Distress				
Altman's Z-Score	5.13	5.14	9.33	9.54
IV. Liquidity				
Quick Assets / Current Liabilities	2.19	2.78	2.26	4.33
Current Assets / Current Liabilities	3.14	3.89	2.30	4.77
Logarithm of Total Assets	8.50	8.31	0.92	0.77
IV. Receivables				
Accounts Receivable / Current Year Accounts Receivable + Prior Year Accounts Receivable	0.55	0.53	0.12	0.09
Current Year Accounts Receivable / Total Sales	0.19	0.17	0.16	0.06
V. Inventory				
Inventory / Sales	0.74	0.17	2.49	0.18
Inventory / Total Assets	0.19	0.17	0.18	0.15
Inventory / Current Assets	0.55	0.31	1.42	0.23
Gross Profit / Total Assets	0.32	0.35	0.38	0.21
Sales – (Net Sales – Cost of Sales)	$802,627,571	$940,448,184	$1,244,983,372	$2,511,777,541
Gross Margin / Two-Year Gross Margin	0.91	1.10	0.46	0.61
VI. Profitability				
Sales / Total Assets	0.83	1.06	0.61	0.61
Earnings before Interest and Taxes	$ 5,742,494	$236,615,987	$ 102,577,279	$ 637,957,817
Net Profit / Total Assets	-0.32	-0.06	1.28	0.41
Net Profit / Total Sales	-5.55	-0.05	22.44	0.37
Operating Income / Total Income				
VII. Additional				
Net Fixed Assets / Total Assets	0.14	0.22	0.14	0.22
Cash / Total Assets	0.13	0.14	0.15	0.14
Working Capital / Total Assets	0.37	0.49	0.29	0.60

Linguistic variables. I utilized the Linguistic Inquiry and Word Count (LIWC) version 2015 software package to perform a linguistic analysis on the dataset. Utilizing a master dictionary of over 6,000 words and emoticons, LIWC works by analyzing the entire text and then reporting the percentage of total words that match each proprietary dictionary. Many dictionaries include easily identifiable words, such as nouns and adjectives, while a number of other

dictionaries include qualitatively defined word categories such as emotional content, spatial orientation and references to status.

Due to the importance of correctly identifying qualitative word categories, the software developers go through a multi-step process for developing these dictionaries. This includes first generating a list of all possible words from dictionaries and thesauruses that relate to the particular dictionary, then selecting only those words that are agreed on by a panel of judges. The software developers then compare the internal validity for each word and retain only those words that are related to other words within the same dictionary in an expectable way (Pennebaker Conglomerates, Inc, 2015).

In their study on internal validity, Kahn, Tobin, Massey & Anderson (2007) performed a series of three experiments testing if LIWC was able to correctly identify positive and negative emotional content utilizing the Emotion Dictionary on written autobiographical memories and films that provoke emotional reactions. The study found that all three experiments supported the construct validity of the LIWC dictionary and that the reported emotional content not only had a high internal consistency but also appeared to measure what it was supposed to measure based on participant feedback.

Prior literature and expectations. There are a number of linguistic "red flags" that have been identified by previous studies to predict deception in general. Of particular importance were affect (e.g. higher percentage of negative emotions), non-immediacy (e.g. past or future tense used to distance the writer from the topic), less personal pronouns and more impersonal pronouns, quantity (e.g. a higher word count when compared to similar descriptions), uncertainty (e.g. use of passive verb tense), a lack of detail, complexity, diversity, expressivity, and specificity (Zhou, Burgoon, Nunamaker, & Twitchell, 2004; Fuller, Biros, Twitchell, Burgoon,

& Adkins, 2006; Hancock, Curry, Goorha, & Woodworth, 2008; Humphreys, Moffitt, Burns, Burgoon, & Felix, 2011; Hauch, Blandon-Gitlin, Masip, & Sporer, 2015).

The current study utilized six categories of words to measure the above signs of deception based on findings from previous literature: Quantity, Complexity, Uncertainty, Expressivity, Specificity, and Non-Immediacy. By utilizing a number of different measures for each category, the study was able to test variables related to those found in prior literature that may show a significant difference between fraudulent and non-fraudulent financial descriptions. For example, even though previous literature illustrates that negative emotions are a predictor of deception, a significant lack of positive emotions may provide a better prediction of fraudulent financial statements. Therefore, the study tested both negative and positive emotions as variables for building the predictive models.

Of particular interest are the five variables Achievement, Power, Reward, Risk, and Money. These are new variables that were included within the most recent version of the LIWC software published in 2015. Due to their recent addition, the literature appears relatively sparse and no articles were found that specifically utilize these variables or measure their internal validity. However, the dictionaries for each variable in the 2015 edition were selected in the same manner as previous dictionaries within the LIWC software, therefore one can reasonably expect the same results on internal consistency that were obtained in the 2007 study by Kahn, Tobin, Massey & Anderson.

I included the variables Achievement, Power, Reward, Risk, and Money due to the expectation that there may be significant differences in the way these word categories are utilized for fraudulent and non-fraudulent financial reporting, even though I found no prior literature utilizing these specific word categories for fraud detection. With the above information in mind,

Table 3 displays the seven categories of 35 linguistic variables used in the study, including descriptive statistics and positive or negative expectations for each variable.

Table 3

Linguistic Variables with Descriptive Statistics

Dimension	Description	Expectation	FFS		N-FFS	
I. Quantity			***M***	***SD***	***M***	***SD***
Word quantity	# of words	+	10,584	7,773	7,812	2,164
Verbs	# Verbs / # Words	-	5.94	0.78	5.89	1.12
% Modifiers	% of adjectives and adverbs	+	4.89	0.59	4.80	0.63
Functionality - S	# function words / # of sentences	+	9.56	2.37	8.99	1.89
Functionality - W	# function words / # words	+	37.4	1.88	36.65	3.16
II. Complexity						
Sentence Length	# words / # sentences	+	28.31	3.69	27.31	3.66
Large Words	% Words longer than six letters	+	33.65	2.90	33.18	2.45
Pausality	# punctuation marks / # sentences	+	3.78	1.89	3.74	1.48
Differentiation	% words that make a distinction	+	1.96	0.39	1.98	0.53
III. Uncertainty						
Tentativeness	% words that express doubt	+	1.97	0.61	1.93	0.71
Certainty	% words that express certainty	-	0.79	0.19	0.77	0.22
IV. Expressivity						
Negations	% words that express negations	+	0.53	0.15	0.56	0.21
Positive Emotions	% words that express positive emotions	-	2.53	0.37	2.44	0.49
Negative emotions	% words that express negative emotions	+	0.98	0.29	1.13	0.28
Unique words	# unique words / # of words	-	73.25	2.26	71.83	4.09
Quantifiers	# words that express quantity	-	1.82	0.47	1.73	0.26
V. Specificity						
Perceptual see	% words referring to sight	-	0.13	0.09	0.18	0.10
Perceptual hear	% words referring to hearing	-	0.03	0.03	0.02	0.04
Perceptual feel	% words referring to feeling	-	0.04	0.04	0.06	0.08
Past Focus	% words referring to past tense	+	2.32	0.44	2.22	0.53
Present Focus	% words referring to present tense	-	3.25	0.55	3.16	0.68
Future Focus	% words referring to future tense	+	1.07	0.31	1.07	0.41
VI. Non-Immediacy						
Singular pronoun	% singular pronouns	-	0.01	0.02	0.01	0.01
Pronoun	% pronouns	-	4.15	1.46	4.13	1.92
We reference	% words referencing “we”	-	1.99	1.56	1.94	1.79
Personal pronoun	% personal pronouns	-	2.11	1.56	2.06	1.85
You	% “you”	+	0.11	0.02	0.02	0.04
She, he	% singular third-person reference	+	0.01	0.01	0.01	0.01
They	% plural third-person reference	+	0.09	0.05	0.08	0.05
Impersonal pronoun	% impersonal pronoun	+	2.04	0.41	2.07	0.48
VII. Business						
Power	% words that reference power	+	2.37	0.46	2.23	0.38
Reward	% words that reference reward	+	0.87	0.20	0.71	0.19
Risk	% words that reference risk	+	1.33	0.44	1.26	0.30
Achievement	% words that reference achievement	+	1.72	0.41	1.60	0.47
Money	% words that reference money	+	7.83	0.96	7.62	0.94

Note: + denotes an expectation of positive impact on odds for FFS, while – denotes and expectation of negative impact on odds for FFS.

Preliminary Variable Reduction

In order to choose only the most important independent variables for the final models, I used a multi-step method of variable reduction incorporating multicollinearity diagnostics, the Akaike Information Criterion (AIC), stepwise regression, and finally forced-entry regression with a comparison of significance values.

Before beginning variable selection, I first addressed any problems of multicollinearity observed between variables by calculating the Variance Inflation Factor (VIF) for each variable in the Financial and Linguistic data sets. Previous literature demonstrates that a VIF of between five and ten appears to be an acceptable cutoff for eliminating variables (Hair, Anderson, Tatham, & Black, 1995; Kennedy, 1992); I chose a VIF of five as the threshold for variable elimination since many financial variables were highly correlated to each other. Choosing a lower threshold further reduces any problems of multicollinearity occurring in the predictive models.

To control for multicollinearity, I first looked at pairs of variables that had a VIF of greater than five and were highly correlated with each other. I then removed one variable from the model that had the lowest significance and re-calculated the VIF for all remaining variables. This procedure was repeated until all variables had a VIF of between one and five, reducing the financial variables from 27 to 15 and the linguistic variables from 33 to 24.

After addressing for multicollinearity, I then used AIC to further reduce the variable set. Calculated as AIC = n ln(SSE)−n ln(n) + 2p, this criterion is used to test each possible model's goodness of fit relative to the overall model complexity. Models with larger amounts of predictor variables are penalized more than models with fewer predictor variables. While this calculation is generally a good indicator of variable selection for complex models, one drawback of AIC is

its tendency to select too many variables when reducing models with a large number of input variables (Spanos, 2010). I accounted for this drawback by using stepwise regression to further reduce the number of input variables based on statistical significance.

After normalizing the dataset, I used the R statistical software package BMA (Raftery, Hoeting, Volinsky, Painter, Yong, 2015) to calculate a best-fit financial model with AIC = 135.69, reducing the number of input variables from 15 to 12. I used the same method to calculate a best-fit linguistic model with AIC = 99.09, reducing the number of input variables from 24 to 17. Finally, I calculated a best-fit combined model utilizing all of the variables with AIC = 63.21, reducing the number of input variables from 39 to 27.

At this stage in variable reduction, I chose two separate methods of determining variable importance due to fundamental differences between Logistic Regression/ANN models and the Random Forest Analysis.

Final Variable Reduction - Logistic Regression and ANN Models

I performed backwards stepwise regression on each of the three variable sets (Financial, Linguistic and Combined) to eliminate any variables that were not significant to $p = .05$. Table 4 illustrates the final models used for each variable set.

Table 4

Full Variable Set for Logistic Regression and ANN Models

Model Type	Sig.	VIF
Financial		
1. Debt / Equity	.005	1.62
2. Sales Growth	.018	1.08
3. Inventory / Total Assets	.003	1.56
4. Gross Profit / Total Assets	.011	2.29
5. Sales / Total Assets	<.001	3.46
6. EBIT	.018	1.33
Linguistic		
1. Word Count	.023	1.45
2. Modifiers	.001	1.37
3. Pausality	.023	1.40
4. Differentiation	.007	3.19
5. Negations	.002	1.78
6. Positive Emotions	.049	1.49
7. Negative Emotions	<.001	1.47
8. Function Words	.019	3.46
9. Perception - See	.023	1.56
10. Singular Pronoun "I"	.025	1.66
11. Pronoun	.002	2.99
12. "They"	.042	1.91
13. Power	.011	1.95
14. Reward	.001	1.35
15. Risk	<.001	1.92
Combined		
1. Inventory / Total Assets	<.001	2.08
2. Sales / Total Assets	<.001	2.28
3. Net Fixed Assets / Total Assets	.022	1.79
4. Word Count	.010	1.37
5. Modifiers	.002	1.51
6. Positive Emotions	.009	1.48
7. Negative Emotions	<.001	1.49
8. Function Words	.019	3.39
9. Perception - See	.002	1.77
10. Perception - Feel	.004	1.39
11. Pronouns	.002	2.53
12. Power	.024	1.82
13. Reward	.001	1.33
14. Risk	<.001	2.15

There appears to be a large discrepancy in the number of predictor variables for each model, particularly the difference between the Financial Model (N = 6) and the Linguistic (N = 15) and Combined (N = 14) Models. In order to account for this, I used forward stepwise regression with an entry of .05 and a stay of .10 which successfully reduced both the Financial and Linguistic models to six variables each. This defined a reduced set of Linguistic and Combined models to better compare with the Financial Model, illustrated in Table 5.

Table 5

Reduced Variable Set for Logistic Regression and ANN Models

Model Type	Sig.	VIF
Financial		
1. Debt / Equity	.005	1.62
2. Sales Growth	.018	1.08
3. Inventory / Total Assets	.003	1.56
4. Gross Profit / Total Assets	.011	2.29
5. Sales / Total Assets	<.001	3.46
6. EBIT	.018	1.33
Linguistic		
1. Word Count	.008	1.11
2. Modifiers	.028	1.07
3. Negative Emotions	<.001	1.24
4. Power	.001	1.27
5. Reward	<.001	1.02
6. Risk	<.001	1.15
Combined		
1. Debt / Equity	.006	1.61
2. EBIT	<.001	1.25
3. Negative Emotions	<.001	1.38
4. Power	<.001	1.19
5. Reward	.003	1.08
6. Risk	<.001	1.39

Final Variable Reduction - Random Forest Analysis

For the Random Forest Analysis, I first used the R Statistical Software to create one decision tree for each model utilizing the variables selected by AIC and based on a training data

set of N = 88 and a testing set of N = 22 observations. I then utilized an R macro written by Dr. Yonggang Lu (personal communication, May 6, 2016) that systematically tested the importance for each variable as it is entered into the decision tree analysis based on GINI and entropy analysis. After generating three lists of variable importance, I then discarded any variable that had an importance rating of less than two standard deviations below the highest rated variable, generating the following three models illustrated in Table 6.

Table 6

Full Variable Set for Random Forest Analysis

Model Type	Importance	VIF
Financial		
1. Net Fixed Assets / Total Assets	20.12	1.64
2. Debt / Equity	16.33	1.06
3. Working Capital / Total Assets	15.67	1.18
4. Sales / Total Assets	13.09	1.59
5. EBIT	8.95	1.32
6. Sales Growth	8.76	1.31
Linguistic		
1. Negative Emotions	20.12	1.39
2. Function Words	17.77	1.17
3. Reward	10.10	1.22
4. Word Count	8.48	1.29
5. Modifiers	8.20	1.17
6. Function Words	5.01	1.89
7. Differentiation	4.59	2.11
8. Perception “See”	4.59	1.32
9. Impersonal “They” pronouns	4.45	1.78
Combined		
1. Sales / Total Assets	14.71	1.19
2. Working Capital / Total Assets	13.81	1.25
3. Altman’s Z-Score	13.10	1.09
4. Gross Margin / 2-Year Gross Margin	12.49	1.11
5. Reward	12.12	1.19
6. Net Fixed Assets / Total Assets	9.98	1.16
7. Negative Emotions	8.48	1.08
8. Word Count	7.41	1.19

There again appears to be a large discrepancy in the number of predictor variables for each model, particularly the difference between the Financial Model (N = 6) and the Linguistic (N = 9) and Combined (N = 8) Models. In order to account for this, I removed the least-important variable until each model had only six variables remaining. This defined a second set of reduced Linguistic and Combined Models to better compare with the Financial Model when running the Random Forest Analysis, illustrated in Table 7.

Table 7

Reduced Variable Set for Random Forest Analysis

Model Type	Importance	VIF
Financial		
1. Net Fixed Assets / Total Assets	20.12	1.64
2. Debt / Equity	16.33	1.06
3. Working Capital / Total Assets	15.67	1.18
4. Sales / Total Assets	13.09	1.59
5. EBIT	8.95	1.32
6. Sales Growth	8.76	1.31
Linguistic		
1. Negative Emotions	20.12	1.31
2. Function Words	17.77	1.04
3. Reward	12.88	1.10
4. Word Count	10.19	1.09
5. Modifiers	8.49	1.13
6. Function Words	8.20	1.42
Combined		
1. Sales / Total Assets	15.09	1.18
2. Working Capital / Total Assets	13.91	1.12
3. Altman's Z-Score	12.56	1.09
4. Gross Margin / 2-Year Gross Margin	10.80	1.08
5. Reward	8.83	1.13
6. Net Fixed Assets / Total Assets	9.76	1.14

Method

I tested the hypotheses by building a total of 15 models; five logistic regression models, five artificial neural networks, and five random forest analyses. I then compared the results of

these models in order to see if combining financial and linguistic input variables had a positive effect on the model's accuracy.

Logistic regression. I used the Statistical Package for the Social Sciences v. 21 (SPSS) to generate five logistic regression models from the final selection of variables utilizing the entry-method with a bootstrap validation of over 10,000 iterations. Fraud was chosen as a selection variable to ensure that each bootstrap iteration had the same number of FFS and N-FFS for proper model validation, and bias corrected accelerated (BCa) statistics were utilized in order to reduce any potential bias from bootstrap validation.

Artificial neural network. Prior research demonstrated no clear way to select the proper architecture for an ANN model, in fact it appears many studies simply did so through trial-and-error. However, there is a large amount of consensus that the two most important factors in neural network architecture are the number of hidden layers and the number of nodes per layer. The majority of neural networks only require one hidden layer; I found this to be true as adding an additional hidden layer decreased the accuracy of the ANN predictions.

In order to determine the proper number of nodes, I created a total of 150 neural networks using 5-fold cross validation with each network ranging from one node to ten. Figure 1 illustrates the average training and testing accuracy graphed for each node; I found that the two types of accuracies came closest to converging when there were six nodes present. Furthermore, the testing accuracy was also greatest at six nodes, and accuracy appeared to systematically decrease when the number of nodes exceeded ten. Based on this result, I chose an architecture with one hidden layer and six nodes for the reduced variable sets. I used the same method to determine the number of nodes for the Linguistic and Combined models utilizing all variables, selecting an architecture of one hidden layer with 11 nodes.

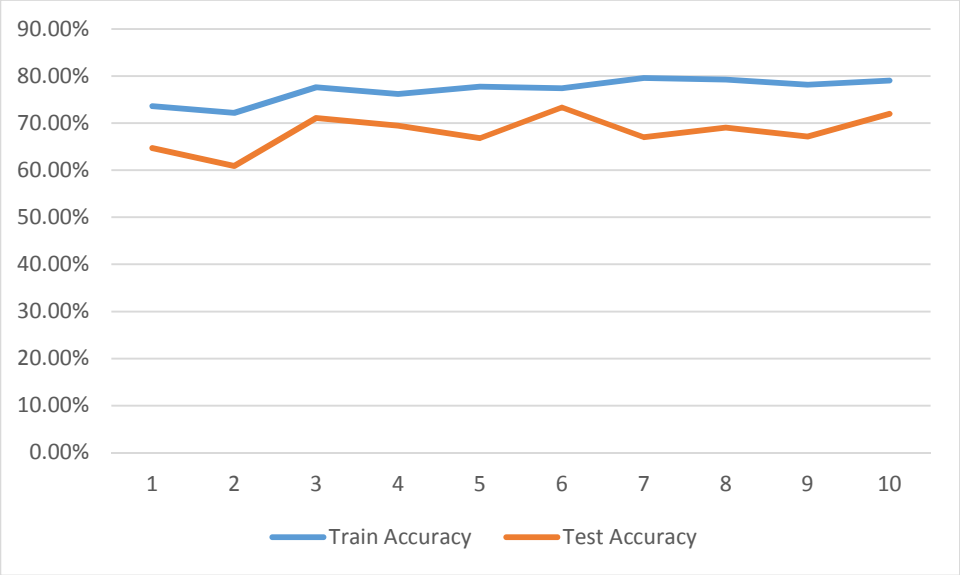


Figure 1: Accuracy for neural networks according to number of nodes.

After determining the correct number of nodes and hidden layers, I used SPSS to create a total of 15 perceptron feed-forward artificial neural networks. For each model, 5-fold cross validation was used with the data randomly partitioned to 60% train, 20% test, and 20% hold-out samples. I ensured that the train, test, and holdout samples were independent from each other for each iteration and that no single observation was used in more than one test or holdout sample. Finally, the results from the five ANN cross-validated samples were averaged to calculate the aggregate accuracy of false negative and false positive error rates for each model.

Random forest analysis. After variable selection, the next step in preparing a random forest analysis is to determine the number of trees where error rates are both lowest and relatively stable. Figure 2 demonstrates that the false negative, false positive and out-of-bag (OOB) error rates appeared to stabilize when the random forest model exceeded 4,000 trees. Therefore, I set each model to include 4,001 trees in total.

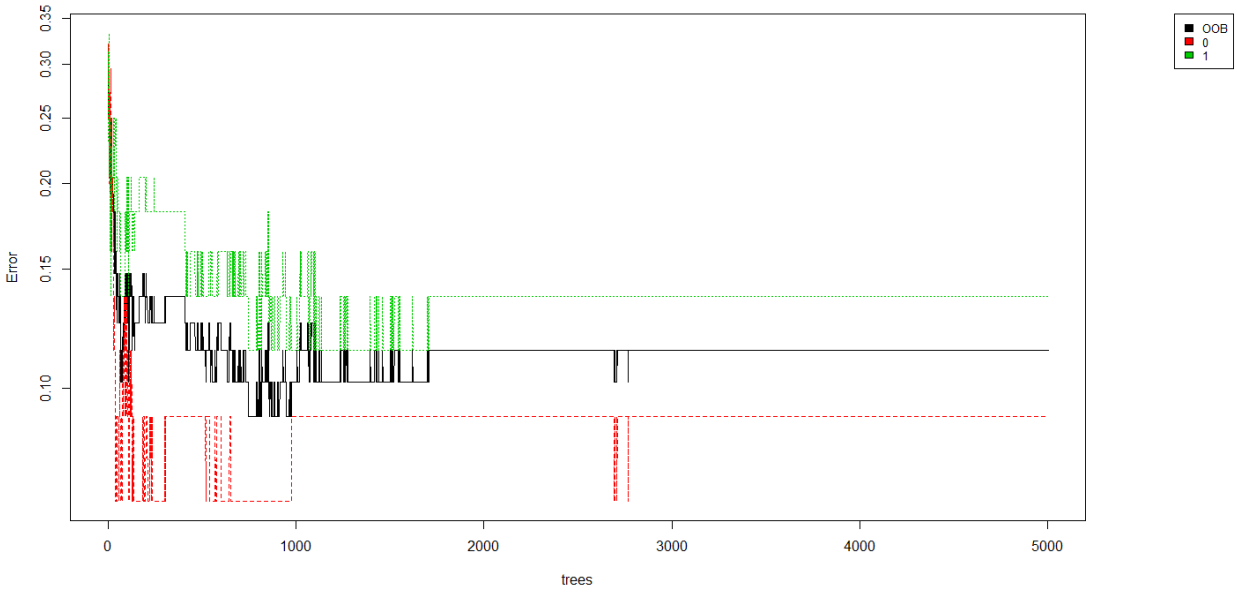


Figure 2: Error rate for random forest models according to number of trees.

After determining the number of trees to use, I created 15 models utilizing 5-fold cross validation with the data randomly partitioned to 80% train and 20% test samples. I ensured that the train and test samples were independent from each other for each iteration and that no single observation was used in more than one test sample. Results from the five cross-validated samples were averaged in order to calculate the aggregate accuracy of false negative and false positive error rates for each model.

Results

Logistic Regression

As seen in Table 8, the results demonstrate that the Combined Regression Model outperforms the Linguistic and Financial Regression Models for overall accuracy when utilizing all predictor variables. However, the Linguistic Regression Model outperforms both the Financial and Combined model when reduced to only six predictor variables. It appears that in both model types, the Combined and Linguistic models report the same false negative accuracy while the Financial Model reports the lowest false negative accuracy.

Table 8

Accuracy of Logistic Regression Models

Model	False Neg. Error[2]	False Neg. Accuracy	False Pos. Error[2]	False Pos. Accuracy	Overall Accuracy
All Variables					
Financial	14	74.54%	15	72.73%	73.64%
Linguistic	5	**90.91%**	6	89.09%	90.00%
Combined	5	**90.91%**	3	**94.54%**	**92.73%**
Reduced to Six Variables					
Financial	14	74.54%	15	72.73%	73.64%
Linguistic	10	**81.81%**	9	**83.63%**	**82.72%**
Combined	10	**81.81%**	10	81.81%	81.81%

As illustrated in Table 9, all three models are statistically significant to $p < .001$, while each individual predictor variable is significant to $p < .05$. The data also demonstrates that the Combined Model provides a more statistically significant prediction according to both the chi-squared and Nagelkerke Pseudo R^2 statistic.

[2] False negative error rate is calculated as the number of times a model predicts an observation is not fraudulent when the observation actually is fraudulent, divided by total number of observations. False positive error rate is calculated as the number of times a model predicts that an observation is fraudulent when the observation is actually non-fraudulent, divided by total number of observations.

Table 9

Summary Results for Logistic Regression Models Utilizing Six Variables

Independent Variable	B	S.E.	Wald	Sig.	Exp(B)
Financial					
1. Debt / Equity	-1.006	.472	7.897	.001	.366
2. Sales Growth	-.467	.204	5.598	.004	.627
3. Inventory / Total Assets	-.462	.210	8.963	.005	.630
4. Gross Profit / Total Assets	-.928	.529	6.546	.011	.395
5. Sales / Total Assets	1.044	.355	12.673	<.001	2.840
6. EBIT	1.397	.652	5.580	.004	4.041
Constant	3.664	2.423	4.306	.027	39.010
Model X^2 42.282 $p < .001$					
Nagelkerke R^2 0.426					
Linguistic					
1. Word Count	-.492	.238	4.641	.008	.611
2. Modifiers	-.281	.156	4.769	.028	.755
3. Negative Emotions	.861	.271	18.242	.000	2.365
4. Power	-.658	.242	12.075	.001	.518
5. Reward	-.725	.219	13.832	.000	.484
6. Risk	-.600	.216	9.986	.000	.549
Constant	6.802	2.272	14.463	.000	899.197
Model X^2 56.334 $p < .001$					
Nagelkerke R^2 0.534					
Combined					
1. Debt / Equity	-1.666	.637	9.685	.002	.189
2. EBIT	2.037	.699	8.704	.006	7.669
3. Negative Emotions	1.059	.366	20.002	<.001	2.882
4. Power	-.803	.287	13.481	<.001	.448
5. Reward	-.813	.243	13.872	<.001	.443
6. Risk	-.644	.271	8.464	.003	.525
Constant	7.460	2.977	8.925	<.001	1736.334
Model X^2 66.128 $p < .001$					
Nagelkerke R^2 .602					

Note: The dependent variable in this analysis is fraud coded so that 0 = fraud and 1 = no fraud. All models were based on a bootstrapped sample of >10,000 iterations with n = 110.

When looking at the Financial Model in particular, it appears that higher Debt / Equity, Sales Growth, Inventory / Total Assets, and Gross Profit / Total Assets increase the likelihood of fraud. On the contrary, a higher reported Sales / Total assets and EBIT tend to decrease the likelihood of fraud. When looking at the odds ratios, EBIT appears to have the greatest effect on

the model with odds increasing by a factor of 4.04 times for every one-unit increase in EBIT on a normalized scale of between one and ten.

Results from the Linguistic Model indicate that a higher word count, usage of modifiers, as well as references to power, risk and reward tend to increase the odds of fraudulent financial statements. Higher scores of Negative Emotion, on the contrary, tend to not only decrease the odds of fraudulent financial reporting, but also appear to have the highest predictive power with odds of fraud decreasing by a factor of 2.37 times for every one-unit increase in Negative Emotion on a normalized scale of between one and ten. This result is contrary to previous literature which demonstrates that higher references of negative emotion tend to increase the odds of deception in general.

Finally, data from the Combined Model demonstrates largely the same results, with higher levels of debt to equity, as well as more references to power, risk and reward increasing the chances for fraudulent financial reporting. Once again, higher reported EBIT and an increased use of negative emotions tend to decrease the odds of FFS. EBIT is once again the most predictive variable in the Combined Model with the odds of FFS increasing by a factor of 7.67 times for every one-unit increase in EBIT on a normalized scale of between one and ten.

As seen in Table 10, results from the Combined and Linguistic Models largely replicate the above patterns, with increased complexity, amount of pause, and references to risk, power, and reward increasing the odds of fraudulent reporting. Interestingly, both negative and positive emotions tend to decrease the odds of fraudulent financial reporting.

Table 10

Results for Linguistic and Combined Logistic Regression Models Utilizing All Variables

Independent Variable		B	S.E.	Wald	Sig.	Exp(B)
Linguistic						
1. Word Count		-.577	.299	3.720	.023	.561
2. Modifiers		-.830	.258	10.332	.001	.436
3. Pausality		-1.091	.470	5.384	.009	.336
4. Differentiation		-.745	.330	5.093	.005	.475
5. Negations		1.354	.453	8.939	.001	3.871
6. Positive Emotions		.630	.309	4.150	.029	1.878
7. Negative Emotions		2.136	.547	15.266	<.001	8.465
8. Function Words		-1.503	.619	5.896	.015	.222
9. Perception - See		.607	.251	5.827	.012	1.834
10. Singular Pronoun "I"		-.575	.273	4.431	.020	.563
11. Pronoun		1.186	.382	9.661	.001	3.275
12. "They"		-.816	.380	4.603	.019	.442
13. Power		-.847	.332	6.514	.003	.429
14. Reward		-1.358	.427	10.129	<.001	.257
15. Risk		-1.467	.433	11.503	<.001	.231
Constant		16.256	5.198	9.779	<.001	11,475,289.714
Model X^2	92.301 $p < .001$					
Nagelkerke R^2	0.762					
Combined						
1. Inventory / Total Assets		-2.237	.608	13.546	<.001	.107
2. Sales / Total Assets		2.095	.583	12.915	<.001	8.129
3. Net Fixed Assets / Total Assets		-.704	.307	5.258	.001	.494
4. Word Count		-1.817	.706	6.630	.002	.162
5. Modifiers		-1.341	.438	9.368	<.001	.262
6. Positive Emotions		1.176	.448	6.904	<.001	3.242
7. Negative Emotions		2.766	.683	16.395	<.001	15.894
8. Function Words		-1.749	.749	5.457	<.001	.174
9. Perception - See		1.063	.347	9.408	<.001	2.896
10. Perception - Feel		2.038	.708	8.294	<.001	7.678
11. Pronoun		1.370	.441	9.643	<.001	3.935
12. Power		-.834	.369	5.099	.002	.434
13. Reward		-1.475	.462	10.210	<.001	.229
14. Risk		-2.190	.621	12.431	<.001	.112
Constant		9.660	5.077	3.621	.018	15,683.183
Model X^2	107.733 $p < .001$					
Nagelkerke R^2	.837					

Note: The dependent variable in this analysis is fraud coded so that 0 = fraud and 1 = no fraud. All models were based on a bootstrapped sample of >10,000 iterations with n = 110.

Artificial Neural Networks

The results in Tables 11 and 12 illustrate that the Combined Model outperforms the Financial and Linguistic Models for both the full and reduced variable sets. Furthermore, the area under curve (AUC) for the receiver operating characteristic (ROC) is highest for the Combined Model in both full and reduced variable sets.

Table 11

Accuracy of Artificial Neural Networks Utilizing All Variables

Variable Type	Validation Set	False Neg. Error	False Neg. Accuracy	False Pos. Error	False Pos. Accuracy	Overall Accuracy
Financial	Train	41	75.15%	39	76.36%	75.76%
ROC 0.836	Test	15	72.73%	12	78.18%	75.45%
	Hold-out	15	**72.73%**	17	69.09%	**70.91%**
Linguistic	Train	4	97.58%	7	95.76%	96.67%
ROC 0.979	Test	9	83.64%	5	90.91%	87.27%
	Hold-out	7	**87.27%**	7	87.27%	**87.27%**
Combined	Train	0	100.00%	2	98.79%	99.39%
ROC 0.988	Test	2	89.09%	4	92.73%	90.1%
	Hold-out	7	**87.27%**	2	94.55%	**90.91%**

Note: Training samples consisted of N = 330 cross-validated observations. Testing and Hold-out samples consisted of N = 110 unique observations. Financial Model utilized six variables, linguistic model utilized 15 variables, Combined Model utilized 14 variables.

Table 12

Accuracy of Artificial Neural Networks Utilizing Six Variables

Model	Validation Set	False Neg. Error	False Neg. Accuracy	False Pos. Error	False Pos. Accuracy	Overall Accuracy
Financial	Train	41	75.15%	39	76.36%	75.76%
ROC 0.836	Test	15	72.73%	12	78.18%	75.45%
	Hold-out	15	72.73%	17	69.09%	**70.91%**
Linguistic	Train	31	81.21%	29	82.42%	81.82%
ROC 0.896	Test	13	76.36%	10	81.82%	79.09%
	Hold-out	17	69.09%	12	78.18%	**73.64%**
Combined	Train	26	84.24%	30	81.82%	83.03%
ROC 0.919	Test	10	81.82%	11	80.00%	80.91%
	Hold-out	13	76.36%	10	81.82%	**79.09%**

Note: Training samples consisted of N = 330 cross-validated observations. Testing and Hold-out samples consisted of N = 110 unique observations. Financial Model utilized six variables, linguistic model utilized 15 variables, Combined Model utilized 14 variables.

Random Forest Analysis

Table 13 illustrates that the Combined Model performs best when utilizing all variables, while the Financial Model performs best when utilizing the reduced set of six variables. The results also demonstrate that while the Combined and Financial model report the same false negative accuracy when utilizing all variables, the Financial Model has a higher false negative and false positive accuracy for the reduced variable set.

Table 13

Accuracy of Random Forest Models

Model Type	False Neg. Error	False Neg. Accuracy	False Pos. Error	False Pos. Accuracy	Overall Accuracy
All Variables					
Financial	8	**85.45%**	11	80.00%	82.73%
Linguistic	9	83.64%	11	80.00%	81.82%
Combined	8	**85.45%**	10	**81.82%**	**83.64%**
6 Variables					
Financial	8	**85.45%**	11	**80.00%**	**82.73%**
Linguistic	15	72.73%	15	72.73%	72.73%
Combined	9	83.64%	12	78.18%	80.91%

Note: All models are based on a random forest analysis of 4,001 trees. 5-fold cross validation was used to independently test the entire data set and obtain accuracy results.

Table 14 lists the six variables used for each model and their respective importance when ranked for that particular model. It appears that some of the results from the logistic regression models are replicated, with Negative Emotions being a top predictor for the Linguistic Model while EBIT and Sales / Total Assets make the top six predictors for both the Financial and Combined models. Taking this into consideration, it also appears that results from the Random Forest Analysis largely replicate results found in previous studies.

Table 14

Importance for Top Six Variables in Random Forest Models

Variable	Fraud	NFraud	Decrease Accuracy	Decrease Gini
Financial Model				
Net Fixed Assets / Total Assets	22.27	38.01	39.27	7.27
Debt / Equity	30.55	26.68	38.59	7.27
Working Capital / Total Assets	29.95	31.14	38.33	6.88
Sales / Total Assets	29.95	25.30	36.50	6.90
EBIT	23.03	30.31	35.64	6.83
Sales Growth	23.22	29.44	33.55	8.30
Linguistic Model				
Negative Emotions	35.11	39.09	48.17	9.92
Function Words	31.82	39.85	46.02	5.87
Reward	23.33	32.29	36.54	8.55
Word Count	9.11	24.99	23.18	7.36
Modifiers	9.80	15.31	16.78	5.82
Function Words	10.46	13.33	16.60	5.94
Combined Model				
Sales / Total Assets	36.90	41.45	50.55	8.69
Working Capital / Total Assets	42.58	36.19	48.72	8.30
Altman's Z-Score	34.63	30.26	41.67	7.60
Gross Margin / 2-Year Gross Margin	32.38	31.90	39.98	4.76
Reward	27.20	25.02	34.20	7.81
Net Fixed Assets / Total Assets	19.24	21.80	27.05	6.31

Discussion

It can be difficult to compare statistical models directly due to the large discrepancy between predictor variables in the Financial Model vs. the Linguistic and Combined Models. As can be seen in Table 15, even though the predictive power of the Combined Model is higher in four out of six statistical tests, both the Financial and Combined models report the highest predictive power in one test each.

That being said, both the Linguistic and Combined Models outperform the Financial Model for logistic regression when all models were reduced to six predictor variables; therefore, it appears that the Financial Model performs the worst out of the three models for logistic regression. When comparing the Linguistic and Combined models utilizing their full variable sets, the Combined Model utilizing 14 predictor variables outperforms the Linguistic Model utilizing 15 predictor variables; thus, it appears that results from logistic regression support the first hypothesis that the Combined Model predicts FFS more accurately overall when compared to the Linguistic or Financial Models alone.

When looking at results from the ANN models, the data illustrates that the Combined Model outperforms all other models for both full and reduced variable sets. Therefore, it appears that data from the ANN models also supports the first hypothesis that the Combined Model predicts FFS better overall when compared to the Linguistic or Financial Models alone.

Finally, data from the random forest analysis demonstrates that while the Combined Model predicts best when utilizing all variables, the Financial Model predicts best when utilizing the reduced set of six variables. This provides mixed results for the hypotheses.

Table 15

Summary of Accuracy Results for All Models

Model Type	False Pos. Error	False Neg. Accuracy	False Neg. Error	False Pos. Accuracy	Overall Accuracy
All Variables					
Logistic Regression					
Financial	14	74.54%	15	72.73%	73.64%
Linguistic	5	**90.91%**	6	89.09%	90.00%
Combined	5	**90.91%**	3	**94.54%**	**92.73%**
Artificial Neural Network					
Financial	15	72.73%	17	69.09%	70.91%
Linguistic	7	**87.27%**	7	87.27%	87.27%
Combined	7	**87.27%**	2	**94.55%**	**90.91%**
Random Forest Analysis					
Financial	8	85.45%	11	80.00%	82.73%
Linguistic	9	83.64%	11	80.00%	81.82%
Combined	8	**85.45%**	10	**81.82%**	**83.64%**
Reduced to Six Variables					
Logistic Regression					
Financial	14	74.54%	15	72.73%	73.64%
Linguistic	10	**81.81%**	9	**83.63%**	**82.72%**
Combined	10	**81.81%**	10	81.81%	81.81%
Artificial Neural Network					
Financial	15	72.73%	17	69.09%	70.91%
Linguistic	17	69.09%	12	78.18%	73.64%
Combined	13	**76.36%**	10	**81.82%**	**79.09%**
Random Forest Analysis					
Financial	8	**85.45%**	11	**80.00%**	**82.73%**
Linguistic	15	72.73%	15	72.73%	72.73%
Combined	9	83.64%	12	78.18%	80.91%

In order to shed more light on these conclusions, I look at a number of statistics that are well-known for calculating the aggregate predictive power of a model while also providing a penalty to models that become overly complex. Table 16 illustrates the AIC, the Bayesian Information Criterion (BIC), and the ROC-AUC for logistic regression models. The table illustrates that the Combined Model has the lowest AIC and BIC for both reduced and full variable sets, while also exhibiting the largest ROC-AUC. These findings also support the

hypothesis that the Combined Model appears capable of predicting better than the Financial or Linguistic Models, even when taking into consideration differences in number of predictor variables for each model.

Table 16

AIC, BIC and ROC for Regression and ANN Utilizing All Variables

Model	AIC	BIC	ROC - AUC
Logistic Regression: All Variables			
Financial	124.21	143.11	0.822
Linguistic	90.79	133.89	0.959
Combined	**73.364**	**113.73**	**0.977**
Logistic Regression: Six Variables			
Financial	124.21	143.11	0.822
Linguistic	110.16	129.06	0.896
Combined	**100.36**	**119.27**	**0.919**

When looking specifically at the false negative error rate in the six statistical tests performed, the combined model has the lowest false negative error rate in two cases, tied for the lowest error rate in three cases, and had the second-lowest error rate in one case. Based on these results, there does not appear to be any evidence supporting the second hypothesis that the Combined Model appears to predict better than other models in regards to the false negative error rate.

When taking a closer look at the predictive power for individual variables in each model, the results appear to confirm most of the data found in previous literature (e.g. Kaminski, Wetzel, & Guan, 2004; Clifton & Phua, 2010; Liu, Chan, Kazmi, & Fu, 2015). When looking at the financial variables, it appears that higher reported solvency, assets and lower amounts of debt tend to decrease the odds of FFS. It is interesting to note that a larger reported Profit / Total Assets tends to increase the odds of FFS, while a larger reported Sales / Total Assets tends to decrease the odds of FFS; further research could determine if there is some sort of interaction

effect between these two specific variables and if they are capable of predicting other important financial aspects such as earnings management.

When looking at the linguistic variables, it appears that higher amounts of complexity such as an increased word count, modifiers, and amount of pause appears to increase the odds of FFS, confirming previous literature on linguistic indicators of deception (Hauch, Blandon-Gitlin, Masip, & Sporer, 2015). Interestingly, while prior research generally shows that decreased amounts of personal pronouns ("I") and increased amounts of third-person pronouns (e.g. "they") tend to increase the chances of deception (Hancock, Curry, Goorha, & Woodworth, 2008), data from the current study suggests that both first-person and third-person pronouns tend to increase the chances of FFS.

Previous literature on linguistic deception also reports that a higher usage of negative emotions in narration appears to increase the odds of deception (Newman, Pennebaker, Berry, & Richards, 2003), whereas the current study illustrates that a higher usage of both negative and positive emotions in Part 7 and Part 7A of the Annual Report appears to decrease the odds of FFS. More research is needed to first replicate the results and second understand why increased emotional expression (both positive and negative) in the annual report tends to decrease the chances of FFS. In either case, results from the current study illustrate that there are important differences between linguistic predictors of deception in general and linguistic predictors specific to FFS that could be addressed in future research.

Another interesting pattern found in the current study is the importance of the linguistic variables Risk, Power and Reward in predicting FFS. All three of these indicators were present in the final logistic regression and artificial neural network models, while Reward also made the top six most important variables for both the Linguistic and Combined random forest analyses.

Since it appears that these specific linguistic indicators have not been used in previous literature to detect FFS, further research is needed to replicate the results. That being said, results from the current study support the notion that references of Risk, Power, and especially Reward appear to be important predictors of FFS.

With the relatively new concept of "big data," many auditors are relying more and more on data analytics in certain stages of the audit, such as testing whether certain accounts require more scrutiny or where to allocate resources. Likewise, with advances in technology such as Enterprise Resource Management Systems and COMPUSTAT, more and more financial data is available at our fingertips. It is no wonder many auditors and financial analysts are turning to this financial data for useful analytics.

Although financial indicators provide a reliable and efficient means of data for analytical purposes, accounting professionals cannot ignore the vast amounts of non-financial data that has also become available through the same increases in computing power. This study gives support to the idea that combining both financial and linguistic predictors could save time and resources during many stages of the audit. This can be particularly useful during the planning stage of an audit when deciding where to allocate important and scarce resources (such as a senior auditing team) to those projects with high indicators of predicted fraudulent activity.

Limitations

There are a number of limitations inherent to this study. I only performed an analysis on financial statements that had accounts receivable and inventory accounts; further research is needed to see if combining financial and linguistic predictors is more accurate for other types of companies, such as financial institutions. Furthermore, I was not able to utilize XBRL data due to the age of the fraudulent cases. The sample included only three financial statements that

contained interactive XBRL data; all of the other cases did not have any XBRL documentation filed with the SEC for that fiscal year. It is possible that XBRL data may be more or less predictive in future models when combined with financial data, something that future studies could focus on.

Another limitation is the fact that the Financial Model contained only six significant predictors, whereas the Linguistic Model contained 15 and the Combined Model contained 14 significant predictor variables. While I attempted to address this problem by creating two sets of models, one with the full variables for each model and one with a reduced six variable set for each model, this does present a problem that could have caused the Financial Model to be less predictive than the Linguistic Model in some cases simply because there were less predictor variables in the Financial Model. Further research could attempt to find models that have the same number of input variables and provide a better comparison between the types of predictors.

There is also a major practical limitation inherent to the specific topic covered in this study. If auditors began using linguistic data within financial statements to predict fraud, it would be fairly easy for a company to run their financial statement through an analysis before submitting it to the SEC and change the wording so there does not appear to be any linguistic indicators of fraud. While this could certainly be the case, the purpose of this study is to perform an exploratory analysis on if financial and linguistic indicators could be combined to provide a more robust predictive model; the implications of this research can be used on more than just the analysis of 10-K financial statements.

Another important limitation is the relatively novel linguistic predictors Power, Risk and Reward, which were added to the LIWC software in 2015. Since so few studies exist that utilize these specific variables from LIWC, the lack of current research and testing for the construct

validity of these particular variables when compared to other variables used from the LIWC software could present a problem in the current study.

A final limitation to the current study is the fact that financial predictors tend to have a significant moderate correlation to each other regardless of how many predictor variables are ruled out. Even though I ensured that no independent variable in either model had a VIF of less than one or greater than four, there were still a number of significant moderate correlations found between the financial predictor variables. This could present a problem when looking at how well the financial predictors work in each model.

Further research could focus on different ways to utilize financial and linguistic data in conjunction to predicting a number of important relationships. An example of a real-life application would be internal auditing, which could utilize an aggregate linguistic analysis of internal e-mails in conjunction with financial indicators to detect if fraud is occurring inside specific departments of a company. Understanding this relationship would require further research on different combinations of linguistic and financial predictors in diverse environments, which could provide better statistical means saving both time and resources during internal and external audits.

References

American Institute of Certified Public Accountants (AICPA). (1988). *The Auditor's Responsibility to Detect and Report Errors and Irregularities.* New York: AICPA.

American Institute of Certified Public Accountants (AICPA). (1997). Consideration of fraud in a financial statement audit. *Statement on Auditing Standards (SAS) No. 82*.

Association of Certified Fraud Examiners, Inc. (2016). *Report to the nations on occupational fraud and abuse.* Retrieved from Association of Certified Fraud Examiners: https://s3-us-west-2.amazonaws.com/acfepublic/2016-report-to-the-nations.pdf

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

Benston, G. J. (2003). The quality of corporate financial statements and their auditors before and after Enron. *Policy Analysis, 12*, 12.

Bump, S. M. (2015). Powering up: How we began with data analytics. *The Journal of Governmental Financial Management, 64*(2), 54-56.

Calderon, T., & Cheh, J. (2002). A roadmap for future neural networks research in auditing and risk assessment. *International Journal of Accounting Information Systems, 3*(4), 203-236.

Chen, S., Yeong-Jia, Goo, J., & Shen, Z.-D. (2014). A hybrid approach of stepwise regression, logistic regression, support vector machine, and decision tree for forecasting fraudulent financial statements. *The Scientific World Journal*.

Clifton Phua, V. L. (2010). *A comprehensive survey of data-mining based fraud detection research.* Monash University, School of Business Systems: Faculty of Information Technology. Clayton: Baycorp Advantage.

Dutta, S. K. (2013). *Statistical Techniques for Forensic Accounting.* New Jersey: FT Press.

Elliot, R., & Willingham, J. (1980). *Management Fraud: Detection and Deterrence.* New York: Petrocelli.

Fraser, L., Hatherly, D., & Lin, K. (1997). An empirical investigation of the use of analytical review by external auditors. *The British Accounting Review, 29*(1), 35-47.

Fuller, C. M., Biros, D. P., Twitchell, D. P., Burgoon, J. K., & Adkins, M. (2006). An analysis of text-based deception detection tools. *AMCIS 2006 Proceedings*, Paper 418.

Green, B. P., & Choi, J. H. (1997). Assessing the risk of management fraud through neural network technology. *Auditing: A Journal of Practice & Theory, 16*(1), 14-28.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate Data Analysis* (3 ed.). New York: Macmillan.

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2008). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes, 45*, 1-23.

Hauch, V., Blandon-Gitlin, I., Masip, J., & Sporer, S. L. (2015). Are computers effictive lie detectors? A meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review, 19*(4), 307-342.

Humphreys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems, 50*(3), 585-594.

Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. *American Journal of Psychology, 120*(2), 263-286.

Kaminski, K., Wetzel, S., & Guan, L. (2004). Can financial ratios detect fraudulent financial reporting? *Managerial Auditing Journal, 19*(1), 15-28.

Kennedy, P. (1992). *A Guide to Econometrics.* Oxford: Blackwell.

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent. *Expert Systems with Applications, 32*, 995-1003.

Kloptchenko, A., Eklund, T., Karlsson, B. B., Vanharanta, C., & Visa, A. (2004). Combining data and text mining techniques for analysing financial reports. *Intelligent Systemes in Accounting, Finance and Management, 12*, 29-41.

Koskivaara, E. (2004). Artificial neural networks in analytical review procedures. *Managerial Auditing Journal, 19*(2), 191-223.

Liu, C., Chan, Y., Kazmi, S. H., & Fu, H. (2015). Financial fraud detection model: Based on random forest. *International Journal of Economics and FInance, 7*(7), 178-188.

Liu, C., Chan, Y., Kazmi, S., & Fu, H. (2015). Financial fraud detection model: Based on random forest. *International Journal of Economics and Finance, 7*(7), 178-188.

Magnusson, C., Arppe, A., Eklund, T., Back, B., Vanharanta, H., & Visa, A. (2005). The language of quarterly reports as an indicator of change in the company's financial status. *Information & Management, 42*, 561-574.

Neil, M., Fenton, N., & Tailor, M. (2005). Using Bayesian Networks to Model Expected and Unexpected Losses. *Risk Analysis, 25*(4).

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*(5), 665-675.

Pearsall, J. (1999). *Oxford Concise English Dictionary, 10th ed.* New York: Oxford University Press.

Pennebaker Conglomerates, Inc. (2015). *How It Works*. Retrieved from Linguistic Inquiry and Word Count: http://liwc.wpengine.com/how-it-works/

Pershad, R. (2000). A bayesian belief network for corporate credit risk assessment. *Centre for Management of Technology and Entrepreneurship: Faculty of Applied Science and Engineering*.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2005). A comprehensive survey of data mining-based fraud detection research. *Artificial Intelligence Review*, 1-14.

PKF Littlejohn LLP. (2015). *The financial cost of fraud 2015: What the latest data from around the world shows us.* Retrieved from PKF Littlejohn LLP.

Raftery, A., Hoeting, J., Volinsky, C., Painter, I., & Yong, K. Y. (2015). *BMA: Bayesian Model Averaging.* Retrieved from R package version 3.18.6: https://CRAN.R-project.org/package=BMA

Ravinskar, P., Ravi, V., Rao, G. R., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems, 50*, 491-50.

Spanos, A. (2010). Akaike-type criteria and the reliability of inference: Model selection versus statistical model specification. *Journal of Econometrics, 158*(2), 204-220.

Spathis, C., Doumpos, M., & Zopounidis, C. (2002). Detecting falsified financial statements: A comparitive study using multicriteria analysis and multivariate statistical techniques. *The European Accounting Review, 11*(3), 509-535.

Tabar, R. H., & WIllis, J. T. (1985). Empirical evidence on the changing role of analytical review procedures. *Auditing: A Journal of Theory & Practice, 4*(2), 93-109.

Team, R. C. (2016). *R: A language and environment for statistical.* Retrieved from R Foundation for Statistical Computing, Vienna, Austria: http://www.R-project.org/

Uzun, H., Szewczyk, S., & Varma, R. (2004). Board Composition and Corporate Fraud. *Financial Analysts Journal, 60*(3), 33-43.

Wells, J. (2007). *Occupational Fraud and Abuse.* Austin, Texas: Obsidian Publishing.

Zhou, ,. L., Nunamaker, J. F., & Twitchell, D. (2004). Automated linguistics based cues for detecting deception in text-based asynchronous computer-mediated communication: An empirical investigation. *Group Decision and Negotiation, 13*(1), 81-106.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation, 13*, 81-106.

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., & Jr, J. E. (2004). A comparison of classification methods for predicting deception in computer-mediated communications. *Journal of Management Information Systems, 20*(4), 139-165.

Zhou, W., & Kapoor, G. (2011). Detecting evolutionary financial statement fraud. *Decision Support Systems, 50*, 570-575.

Zopounidis, C., & Doumpos, M. (1999). A multicriteria decision aid methodology for sorting decision problems: The case of financial distress. *Computational Economics, 14*(3), 197-218.