Discriminating Fraudulent Financial Statements by Identifying Linguistic Hedging

Sean L. Humpherys
University of Arizona, shumpherys@cmi.arizona.edu

Follow this and additional works at: http://aisel.aisnet.org/amcis2009

Recommended Citation
http://aisel.aisnet.org/amcis2009/400

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 2009 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
Discriminating Fraudulent Financial Statements by Identifying Linguistic Hedging

Sean L. Humpherys
Center for the Management of Information,
University of Arizona, Tucson, Arizona
shumpherys@cmi.arizona.edu

ABSTRACT

Managerial financial fraud is estimated in the billions of dollars annually in the United States. Since fraud includes obfuscation, misdirection, and fabrication, this study proposes using deception theory as a means of detecting fraud in textual portions of financial statements (10K). A corpus of 101 fraudulent 10Ks was collected from the Securities and Exchange Commission along with 101 matching non-fraudulent 10Ks. Natural Language Processing techniques were applied to the corpus to generate raw counts and usage rates of hedging devices: hedging modal verbs, hedging adjectives, hedging adverbs, hedging conjunctions, hedging nouns, and hedging lexical verbs. A classification model, based on logistic regression, successfully discriminates with 69.3% accuracy and accounts for nearly 20% of the observed variance. Two machine-learning algorithms are investigated. Bayesian Network and JRip achieve accuracy results of 62.4% and 67.8% respectively. Both results are better than chance or of human deception detection suggesting the possibility of a diagnostic tool for auditors.

Keywords

Managerial financial fraud, hedging devices, auditing

INTRODUCTION

Fraud refers to the “intentional act to obtain an unjust advantage” (International Federation of Accountants, 2008) and includes a scheme designed to deceive the public (Wallace, 1995). Deception is the intentional transmission of information intended to foster a false conclusion in the receiver (Buller and Burgoon, 1996; Burgoon and Nunamaker, 2004; Carlson, George, Burgoon, Adkins, and White, 2004; Burgoon and Qin, 2006) and therefore is related to fraud. Management financial fraud (MFF) is a type of strategic deception where stakeholders are negatively affected through misleading financial statements (Elliot, 1980).

One of the roles of financial auditors is to certify the authenticity of a financial statement. If there are discrepancies, intentional or unintentional, the auditor has an obligation to discover and communicate them. Yet, even trained auditors are credited with only discovering 12% of fraud (KPMG, 2003). Humans, as lie detectors, are no better at detecting deception, as studies demonstrate that humans are barely better than chance at detecting deception in communication (Bond and DePaulo, 2006). These two facts help explain why management fraud is estimated to be in the billions of dollars annually in the United States (Well, 1997). Because of the societal costs associated with MFF, the American Institute of Certified Public Accountants’ Auditing Standards Board (ASB) released in 2002 a Statement on Auditing Standards (SAS) 99. Under SAS 99, auditors are required to undertake auditing procedures to detect MFF. SAS 99 underscores the need to develop procedures, methods, and tools to assist in countering MFF. Developing an IT artifact that classifies financial statements as potentially fraudulent or not can help auditors meet the SAS 99 guidelines and focus their auditing efforts. This paper explores the developing a reliable IT artifact usable by auditors.

To successfully counter fraud, it must first be detected. Traditional forms of detecting financial manipulation have been too focused on the numerical credibility of key financial indicators. Traditional approaches have ignored the deceptive communication that accompanies the fraudulent financial numbers. Yet, several studies have used linguistic techniques to discriminate deception in interviews, emails, and statements to the police (Zhou, Burgoon, Nunamaker, Twitchell, 2004A and 2004B, Qin and Burgoon, 2007; Fuller, Biros, and Wilson, 2008). This study proposes using deception theory with linguistic theory to discriminate fraudulent financial statements from non-fraudulent financial statements.
The paper is organized as follows. First I describe the type of financial statements used to investigate fraud and deception. I then explain the communicative purposes of hedging as a linguistic device. I discuss a theory of deception to explain how deceivers may use hedging to falsely project an appearance of credibility and to establish a barrier of protection against failed predictions. The research questions will be articulated, hypotheses presented, and a description of the research methodologies employed discussed. Finally the results are presented and discussed.

10K FINANCIAL STATEMENTS AND FRAUD

Annual reports have been used to understand how management communicates with shareholders and with the public. Bettman and Weitz (1983) analyzed letter to shareholders in 181 annual reports finding that unfavorable outcomes were attributed to external causes supposed beyond the control of management. Self-serving patterns of attribution were also found. Elsbach and Sutton (1992) investigated how unlawful acts by management are communicated. Drawing on the theoretical principles of institutional impression management, they found that spokespersons’ tactics shifted attention away from the unlawful act and towards the socially desirable goals of the institution. Abrahamson and Park (1994) investigated the theoretical principles of institutional impression management, they found that spokespersons’ tactics shifted attention away from the unlawful act and towards the socially desirable goals of the institution. Abrahamson and Park (1994) investigated how unlawful acts by management are communicated. Drawing on the theoretical principles of institutional impression management, they found that spokespersons’ tactics shifted attention away from the unlawful act and towards the socially desirable goals of the institution.

Publicly financed corporations in the United States are required to submit financial documents to the U.S. Securities and Exchange Commission (SEC) quarterly and annually. The mission of the SEC is “to protect investors, maintain fair, orderly, and efficient markets, and facilitate capital formation” (SEC, 2008A). One of the required financial statements submitted to the SEC annually is the 10K form. The 10K is submitted electronically and contains several sections where key financial indicators are presented along with a managerial discussion regarding the financial health and future outlook of the corporation.

If the SEC discovers financial manipulation or obfuscation, an Accounting and Auditing Enforcement Release (AAER) is issued against the company and/or against specific stewards of the company. AAERs communicate to the public concerning “civil lawsuits brought by the Commission in federal court and notices and orders concerning the institutions and/or settlement of administrative proceedings” (SEC, 2008B). Subsequent action is taken in the Federal Court for reparations, if it is not settled out of court by the parties. The issuance of an AAER is an indicator of purposeful, not just accidental, earnings manipulation and useful for researching MFF (Dechow, 1996).

Of research interest for this study is the section of the 10K called the Management’s Discussion and Analysis (MD&A) which discusses the financial condition, results of operations, and future outlook of the company. This section is largely textual and subjective. Resulting from Sarbanes-Oxley Act of 2002 (SOX), management, particularly the CEO and CFO, are responsible for creating the corporate ethical culture and accountable for discovering and preventing MFF. The MD&A section provides a unique opportunity to investigate how management crafts language that contains obfuscation, misdirection, hedging, equivocation, deception and fraud.

The SEC makes available 10Ks and AAERs through EDGAR, an online database of SEC forms (http://www.sec.gov/edgar.shtml). 10Ks from 202 publicly traded U.S. companies were randomly selected for analysis. Half the companies have AAERs against them indicating purposeful financial statement manipulation (labeled fraud group) and half have never had an AAER issued against them (labeled the control group). The 101 fraudulent 10Ks were identified by searching randomly for AAERs that included the term ‘10K’. Companies named in AAERs are assumed to be guilty of earnings manipulations (Dechow, 1996). Because of the random nature of the selection of the fraudulent 10Ks, sixty-eight unique SIC codes comprise the sample and the year of 10K publications ranges from 1995 through 2004. For the control group, 101 analogous non-fraudulent 10Ks were chosen by first filtering the list of public companies with Standard Industrial Classification (SIC) codes that exactly matched the companies in the fraud group, a search feature available on EDGAR. This process resulted in dozens to hundreds of potential control companies per SIC code. From within each SIC code group, control companies were randomly selected for a total of 101 non-fraudulent companies. The non-fraudulent companies were confirmed to have no AAERs attached to them, by searching SEC’s list of AAERs, which suggests a history of compliance to SEC regulations. Next, each control group’s 10K was selected to match the same year (or as closely as possible) to the matching 10K’s year from the fraud group. The purposes of these criteria are to minimize potential confounds because of differing economic conditions or differences from disassociated industries when comparing the fraudulent to the non-fraudulent. Since the 10Ks are text-based, natural language processing methods can be used to investigate the linguistic difference between the two groups. Of particular interest is their use of hedging language.
USE OF HEDGING LANGUAGE

Lakoff (1972) defined hedging as “words whose job it is to make things more or less fuzzy.” Hedges are linguistic devices, such as perhaps, might, maybe, etc., used to communicate the speaker’s degree of confidence, or lack of confidence, in a statement (Hyland, 1998; Coates, 1987). Hedging expresses tentativeness or probability and are statements used to avoid commitment (Hyland, 1998).

Consider these two statements which include epistemic words.

The company will experience record profits this year. [1]

The company might experience some profits this year. [2]

It is obvious that the first sentence communicates confidence in the speaker’s opinion and prediction. The second qualifies the prediction with the word might, which communicates uncertain probability, and with some, which communicates an uncertain quantity. In the second sentence the author communicates tentativeness, e.g. hedging, and at the same time protects himself/herself from accountability should the prediction fail.

Through hedging the author reduces the strength of a statement (Zuck and Zuck, 1986). Hedges are any manipulative, non-direct sentence to strategically say less than one means (Markkanen and Schroder, 1989). With hedging the author purposefully makes a lack of explicit commitment to the veracity of a statement (Hyland, 1998). The salient point in these definitions is the strategic nature of hedging. It is a linguistic device chosen with purpose for its communicative powers, which becomes important later as we discuss the strategic nature of deception.

In regards to financial statements, the writers of 10Ks purposefully identify for the reader the uncertain nature of their discussion. Typically, 10Ks include a “Safe Harbor” which purposefully communicates that information in the MD&A contains uncertainty. The writer informs the reader how to identify the uncertainty by identifying a list of common terms used in forward-looking statements, terms such as believes, expects, may, should, etc. Most of the terms are also hedging words. However, the presences of hedging does not by itself signify deception. The practice of using forward-looking statements and hedging words in financial statements is encouraged to promote transparency, according to U.S. law in section 27A of the Securities Act of 1933, section 21E of the Securities Exchange Act of 1934 and the Private Securities Litigation Reform Act of 1995. The first research question is the following:

RQ#1: Do writers of fraudulent financial statements use hedging devices more or less frequently than non-fraudulent ones?

To explore this question, deception needs to be discussed.

DECEPTION THEORY

Theories on deception predict a phenomenon of leakage. Leakages are behavior cues that indicate deceptive communication because of the deceiver’s inability to match behavior cues during non-deceptive communication (Ekman and Friesen, 1969). Interpersonal Deception Theory (IDT) proposes that leakages are non-strategic behavior cues (Burgoon and Burgoon, 1996, Burgoon and Nunamaker, 2004). The unintentional nature of non-strategic leakage results from an inability to behave during deceptive communication exactly as one would normally behave during non-deceptive communication. Because deceptive communication is more complex than non-deceptive communication (Zuckerman, DePaulo, and Rothenthal, 1981, Vrij, Roberts, and Bull, 2001; Qin and Burgoon, 2007) it is thought that cognitive loads in deceptive communication are higher than normal. IDT also predicts strategic behaviors which are purposefully chosen by the deceiver. Because deception is goal-driven, deceivers must “be concerned about appearing credible, allaying receiver suspicions, minimizing their responsibility of deceit, and avoiding unpleasant consequences if deception is detected” (Burler and Burgoon, 1996, pg 216).

PROPOSITION: Deceivers strategically use hedging devices for information management purposes and to avoid responsibility; and they do so more frequently than truth tellers because of the need to obfuscate, misdirect, or fabricate.

The justification for this proposition is that deceivers need to appear credible and hedging increases the perception of credibility (Hyland, 1998) by acknowledging alternative possible outcomes, while affording other benefits to the deception at the same time. Deceivers need linguistic devices to manipulate information along dimensions of completeness and truthfulness, i.e. information management (Burler and Burgoon, 1996). Hedging communicates without completeness and reduces certainty. Hedging creates fuzziness, which can be advantageous to the deceiver.

Additionally, hedging minimizes the contrarian’s ability to hold accountable the sender of a deceptive message. Consider again example sentence #2. If profits decreased instead of increased, the hedger could not be held accountable for any error because the hedger did not take a declarative position. Hedging creates a barrier of protection which the deceiver can hide.
behind, which is a motivational factor of deceivers (Buller and Burgoon, 1996). Since deceivers in MFF have inside knowledge that would negatively affect the company, personal profit, or their self-image if disclosed, the deceiver has a need to obfuscate, misdirect, or fabricate. For these reasons, combined with the deceiver’s goal and need to hide deception, I hypothesize that the use of hedging is a strategic linguistic device used by deceivers more frequently than truth tellers.

H1: Deceivers use hedging devices in fraudulent 10Ks more frequently than in the control group.

The method used to inquire about deception and hedging in financial statements is to use natural language processing techniques. A natural language tool kit provides algorithms written in Python that can parse sentences and words. Lexicons were created that include hedging terms, the creation of which was guided by linguistic research.

RQ#2: Which lexicons of hedging can be used for deception detection?

Although hedging has not been researched in context of deception or MFF, hedging has been well researched by sociolinguists. Hyland identifies various surface features of hedging by identifying modal verbs, adverbs, adjectives, lexical verbs, nouns, and phrases that have hedging uses (Hyland, 1998).

HEDGING MODAL VERBS

The first hedging category is that of modal verbs, although not all modal verbs have hedging purposes. The following are epistemic modal verbs that communicate something less than certainty, therefore useful as hedgers (Hyland, 1998); ought, should, would, could, may, might. These words communicate possibility and probability but do not communicate certainty as does will, must, and won’t.

Some modal verbs are polysemous (Huddleston, 1971), which can create difficulties in using computer automated means for identifying a modal verb with hedging purpose. For the purposes of this study, should will not be included in the lexicon of hedging modal verbs. I sampled 50 uses of should from the 345 uses in the corpus, 9 uses are of a hedging nature and 41 are deontic or have non-hedging uses. May has an alternative meaning of “permission”, but this meaning is rare (Hyland, 1998) and of 50 samples of may in the corpus, all 50 were of the epistemic variety. Thus, may is included among the hedging lexicons. The word can has multiple purposes but no epistemic meaning (Hyland, 1998). Because can is not used to reduce certainty (Hyland, 1998), it is not included in the lexicon of hedging modal verbs. However, when combined with other hedgers can may be part of a larger hedging clause, such as in the clause “there can be no assurance.” These sentences (Examples 3 and 4) from the corpus communicate hedging.

However, there can be no assurance that this investment will be completed. [3]

There can be no assurance that any of these companies will be successful or achieve profitability or that we will ever realize a return on our investments. [4]

In these examples the other hedging words, e.g. however and assurance, are required for the sentence to communicate epistemic uncertainty. These examples imply “let the buyer beware” and “uncertainty in our claim exist.” Phrases and clauses with hedging purposes were not included in this study but are worthy of investigation at a future date. Since these examples include other hedging words which are included in my hedging lexicons, these two example sentences will be counted in the hedging analysis. All these facts lead to H2.

H2: Deceivers will use more hedging modal verbs and fewer certainty modal verbs than the control group in 10Ks.

Apart from modal verbs, there are other epistemic lexical verbs which can be used as hedgers.

According to Hyland, epistemic lexical verbs are “the most transparent means of coding the subjectivity of the epistemic source and are generally used to hedge either commitment or assertiveness” (Hyland, 1998, pg.119). These lexical verbs include the following:

Table 1: Hedging Lexical Verbs

<table>
<thead>
<tr>
<th>anticipat*</th>
<th>expect*</th>
<th>indicated</th>
<th>propos*</th>
</tr>
</thead>
<tbody>
<tr>
<td>appear*</td>
<td>feel*</td>
<td>indicating</td>
<td>rekon*</td>
</tr>
<tr>
<td>assum*</td>
<td>Felt</td>
<td>infer*</td>
<td>seek*</td>
</tr>
<tr>
<td>assur*</td>
<td>Forecast*</td>
<td>intend*</td>
<td>seem*</td>
</tr>
<tr>
<td>believ*</td>
<td>guess*</td>
<td>plan*</td>
<td>show*</td>
</tr>
</tbody>
</table>
HEDGING CONJUNCTIONS, ADJECTIVES, ADVERBS AND NOUNS

Conjunctions, such as although, serve the linguistic purpose of epistemically countering a previous statement. Although communicates to the user that there are exceptions, violations, or uncertainty regarding the primary claim. Other conjunctions that can serve the same purpose include if and notwithstanding.

In scientific discourse articles, Hyland found that adjectives, adverbs, and nouns constitute 53% of the lexical hedgers (Hyland, 1998). The following non-exhaustive lists (Table 2) are those words which have been found to be used as hedgers because of their epistemic, less-than-certain meanings.

**Table 2: Other Hedging Lexicons**

<table>
<thead>
<tr>
<th>Hedging Adjectives:</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>about</td>
<td>apparent</td>
<td>approximate</td>
<td>assumptive</td>
</tr>
<tr>
<td>indicative</td>
<td>many</td>
<td>most</td>
<td>predictive</td>
</tr>
<tr>
<td>possible</td>
<td>potential</td>
<td>probable</td>
<td>presumptive</td>
</tr>
<tr>
<td>suggestive</td>
<td>connotative</td>
<td>deductive</td>
<td>speculative</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hedging Adverbs:</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>almost</td>
<td>around</td>
<td>apparently</td>
<td>approximately</td>
</tr>
<tr>
<td>eventually</td>
<td>essentially</td>
<td>generally</td>
<td>however</td>
</tr>
<tr>
<td>likely</td>
<td>likelier</td>
<td>likeliest</td>
<td>maybe</td>
</tr>
<tr>
<td>nearly</td>
<td>normally</td>
<td>occasionally</td>
<td>often</td>
</tr>
<tr>
<td>mostly</td>
<td>partially</td>
<td>perhaps</td>
<td>possibly</td>
</tr>
<tr>
<td>potentially</td>
<td>presumably</td>
<td>probably</td>
<td>quite</td>
</tr>
<tr>
<td>rarely</td>
<td>relatively</td>
<td>slightly</td>
<td>soon</td>
</tr>
<tr>
<td>some</td>
<td>somehow</td>
<td>somewhat</td>
<td>unlikely</td>
</tr>
<tr>
<td>usually</td>
<td>virtually</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hedging Nouns:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>possibility</td>
<td>probability</td>
</tr>
</tbody>
</table>

Some of these words modify numerical data and therefore should be present in 10Ks, e.g. most, some, about, approximately. They can be used by an author to express a quantity, frequency, or probability with inexactness, which is a hedging strategy.

H3: Deceivers use hedging conjunctions, hedging adjectives, hedging adverbs, and hedging nouns at greater frequency in 10Ks than the control group.

One of the purposes of this research is to create automated systems that can discriminate between fraudulent and non-fraudulent financial statements based on cues consistent with deception. Machine learning algorithms using Zhou, Burgoon, Nunamaker, and Twitchell’s 27 linguistic-based cues have been successful in discriminating between deceivers and truth-tellers with 60% to 80% classification accuracies (Zhou et al., 2004B). Cues included word quantity, affect term ratio, passive voice ratio, non-immediacy terms ratio, etc. This fact prompts the following research question.
RQ#3: Can the presences or absence of hedging devices be used as a reliable indicator of deception (or fraud) using classification algorithms?

It is anticipated that a classification algorithm, such as JRIP and Bayesian Network, will be successful at discriminating between fraudulent 10Ks and non-fraudulent 10Ks.

Theories should be proffered prior to any statistics. I proffer an alternative theory which may explain contrary results should the hypotheses prove false. There may be a disadvantage to using hedging which discourages managers from using them. Consider the first two sentences again.

The company will experience record profits this year. [1]

The company might experience some profits this year. [2]

Sentence 1 is the type of claim a competent, successful manager might make or that stakeholders might expect from a competent manager. Sentence 2 could communicate a lack of competency in one’s ability to generate profits. Hedging devices may negatively reflect on one’s image as competent. So the deceiver is left with competing motivations: use hedging to protect oneself and obfuscate damning information or appear competent by avoiding hedging devices.

RESULTS

H1 states that deceivers use hedging devices in fraudulent 10Ks more frequently thantruthtellers. In an independent samples, one-sided t-test, deceivers do use more total hedging words in 10Ks than the control group (t(200)=4.293; p<=0.0005; mean diff.=79.4). However, controlling for document length of the 10K by using a ratio of hedging count to token count, the statistical significance disappears (t(200)=0.125; p=0.900; mean diff.=-1.3). Tokens are words, abbreviations, contractions, numeric entities, and punctuation marks. Therefore, H1 receives mixed support based on whether the research interest is in the writer’s choice to compose and purposefully use more hedging devices than others or the interest is in the usage rate of hedging devices for which there appears to be no difference in rate.

H2 states that deceivers will use more hedging modal verbs and fewer certainty modal verbs than truthtellers in 10Ks. Results are presented in Table 3. Deceivers use could, may, might, and would at a greater rate than the control group. Could is statistically significant in both raw count and usage rate (t(200)=2.292, p=0.023) and is used more by the fraud group than the control group. The certainty modal verbs, may, might, and would are used more frequently by the fraud group, but the rate differences are not statistically significant: may (t(199)=0.846, p=0.399), might (t(124)=0.793, p=0.430), would (t(197)=0.135, p=0.893). Will, a certainty modal, is statistically significant (t(200)=-2.356, p=0.019) and used more frequently by the control group. Ought and won’t are not used by either group in this corpus.

<table>
<thead>
<tr>
<th>Hedging Modal Verbs</th>
<th>Fraud</th>
<th>Control</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ought</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>could (couldn’t)</td>
<td>14.0</td>
<td>9.8</td>
<td>4.2*</td>
</tr>
<tr>
<td>may</td>
<td>23.4</td>
<td>21.1</td>
<td>2.3</td>
</tr>
<tr>
<td>might</td>
<td>1.0</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>would (wouldn’t)</td>
<td>8.4</td>
<td>8.5</td>
<td>-0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Non-Hedging Modal Verbs</th>
<th>Fraud</th>
<th>Control</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>can (cannot, can’t)</td>
<td>7.9</td>
<td>10.2</td>
<td>-1.3</td>
</tr>
<tr>
<td>must (mustn’t)</td>
<td>1.4</td>
<td>1.6</td>
<td>-0.2</td>
</tr>
<tr>
<td>shall</td>
<td>0.2</td>
<td>0.8</td>
<td>-0.6</td>
</tr>
<tr>
<td>should (shouldn’t)</td>
<td>3.0</td>
<td>3.7</td>
<td>-0.7</td>
</tr>
<tr>
<td>won’t</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
H3 states that deceivers use hedging conjunctions, hedging adjectives, hedging adverbs, and hedging nouns at greater rates in 10Ks than the control group. Although a mean difference is observable and in the predicted direction, none of the hedging ratios produce a significant difference in independent samples t-tests (see Table 4).

### Table 4: Usage rate of hedging devices

<table>
<thead>
<tr>
<th></th>
<th>Fraud</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hedging Modal Verbs</td>
<td>46.8</td>
<td>40.0</td>
</tr>
<tr>
<td>Hedging Lexical Verbs</td>
<td>86.7</td>
<td>93.9</td>
</tr>
<tr>
<td>Hedging All Verbs</td>
<td>133.5</td>
<td>133.9</td>
</tr>
<tr>
<td>Hedging Adj</td>
<td>15.6</td>
<td>16.1</td>
</tr>
<tr>
<td>Hedging Adv</td>
<td>50.4</td>
<td>50.4</td>
</tr>
<tr>
<td>Hedging Nouns</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>Hedging Conjunctions</td>
<td>14.1</td>
<td>14.6</td>
</tr>
<tr>
<td>All Hedging Words</td>
<td>213.8</td>
<td>215.2</td>
</tr>
<tr>
<td>Certainty Modal Verbs</td>
<td>36.7</td>
<td>46.2</td>
</tr>
<tr>
<td>All Modal Verbs</td>
<td>86.5</td>
<td>89.9</td>
</tr>
</tbody>
</table>

Mean usage rate per 10,000 tokens; none found to be statistically different

To investigate the predictive ability of the hedging lexicons, I performed a binary logistic regression with likelihood-ratio test (nested model design) using the raw counts of hedging variables and the usage ratio of the raw count to the token count. This later calculation is an interaction under the assumption that document length interacts with the raw counts and will therefore improve the overall logistic model.

Using a nested design requires a complete model and a partial model for comparison. The dependent variable in both is group (fraud or control). The partial model includes the following independent variables: token count, hedging modal verbs count, hedging adjective count, hedging adverb count, hedging noun count, hedging conjunction count, and hedging lexical verb count. The complete model adds the hedging modal verbs ratio (calculated by dividing by token count), hedging adjective ratio, hedging adverb ratio, hedging noun ratio, hedging conjunction ratio, and hedging lexical verb ratio. The ratio variables can be considered interactions between the raw count and the token count. The theory described earlier predicts the importance of the raw count and ratio for each of these variables, including the direction of differences among fraud and control group. Thus, testing with a nested model design ensures one does not commit Type 2 errors of leaving out an important variable, while ensuring that one does not commit a Type 1 error of including irrelevant variables.

The partial model has a -2LL of 252.8 and Nagelkerke R square of 0.168. The complete model has a -2LL of 247.6 and R square of 0.198. The smaller the -2LL, the better the model maximizes likelihood. Nagelkerke R square can be interpreted somewhat similar to a multivariate regression’s R square, the amount of variance accounted for by the independent variables, excepting the values for Nagelkerke’s R square are typically lower and should not be directly compared to the familiar regression R square (Norusis, 2003, pg 334). The omnibus test of model coefficients results in a significant difference between models to a value of 0.002. Therefore I conclude that the raw counts and the ratios variables are required for a complete logistic model of deception in 10Ks. The logistic regression also predicts classification of fraud or control cases. The partial model predicts with 66.8% accuracy. The complete model has a 69.3% accuracy rate, improving on both the classification of fraud (64.4%) and of the control group (74.3%).

Research question 2 prompts an investigation of machine learning algorithms as predictive classifiers. A Bayesian Network employed on the same variables as in the logistic regression results in a 62.4% accuracy classification with a root mean squared error of 0.5145; precision of fraud at 64.4%; and precision of control group 60.9%. A JRip classification algorithm
results in 67.8% accuracy, a root mean squared error of 0.4717, precision of fraud at 70.1%, and precision of control group 65.5%. The JRip implements a propositional rule learner with Repeated Incremental Pruning to Produce Error Reduction (Cohen, 1995). A 10-fold cross validation was used for each test.

**DISCUSSION**

After reviewing deception theory and linguistic theory I posit the following: deceivers strategically use hedging devices for information management purposes and to avoid responsibility, and they do so more frequently than truth tellers because of the need to obfuscate, misdirect, or fabricate. The rationale is that hedging devices provide many of the communicative advantages that deceivers desire from lexical devices. However, this study provides mixed support for the theory but the findings are nonetheless important to understanding how fraudulent 10Ks are crafted.

H1 predicted that deceivers will use more hedging devices in fraudulent 10Ks than truth tellers and have greater usage rates. Deceivers use more hedging devices as a raw count but show no difference in usage rate. One cannot say that deceivers use hedging devices more frequently, but something is certainly influencing writers of fraudulent 10Ks to write more and use more total words, including more hedging devices than in the control group. This last finding is consistent with deception research, that deceivers purposefully use more words in written documents than truth tellers (Zhou et al. 2004A, Zhou et al. 2004B).

The reason for lack of concrete support may be two-fold. First, business and sales language may shunt the general use of hedging devices because of detrimental affects to one’s image of managerial competence. It is plausible that no manager, truthful or deceitful, desires to communicate uncertainty when discussing the outlook of their company, thereby erasing any statistical difference one might find from deception. A comparison with other types of financial documents can aid this inquiry.

H2, which predicted that deceivers will use more hedging modal verbs and fewer certainty modal verbs in 10Ks than the control group, received mixed support. Could is used more frequently by those in the fraud group, but there is no difference in frequency use among the other hedging modal verbs (e.g. ought, may, might, would, should). Even among the certainty modal verbs, only will is used differently. The control group uses will more frequently then the fraud group.

The fraud group seems to prefer could and minimizes the use of will. This observation is in line with the theory because could is less committal than will and being less committal is a deceptive strategy. But other than these two modal verbs there is no discernable difference between hedging modal verbs collectively used by either group. The same reasons cited above may be at play here too.

H3, which predicted that deceivers use hedging conjunctions, hedging adjectives, hedging adverbs, and hedging nouns at greater frequency in 10Ks than the control group, did not receive support. None of these hedging constructs are used more or less frequently by either group. The sum of all hedging devices, as a single latent variable of hedging usage rate, is not used differently by either group. It was observed that the collection of certainty modal verbs is used more frequently by the control group than the fraud group ($t_{200}=-2.284, p=.01$). This last fact coincides with the previous findings regarding the use of will, which is the predominant verb among the certainty modal verbs.

Research question 2 regarding classification of fraud did receive support. The complete logistic regression model classifies with 69.3% accuracy and accounts for 19.8% of the variance in the dependent variable. JRip and Bayesian Network classified fraud or control better than chance and better than human deception detection which averages only 54%. These findings hold promise that a tool can be developed to aid auditors in their assessment of the credibility of financial statements. Further investigation, developmental refinement, and addition of other linguistic cues may results in improved discriminatory power.

**CONCLUSION**

One scientific contribution of this paper is the compilation of a hedging lexicon used in 10Ks, which specifically includes hedging modal verbs, hedging adjectives, hedging adverbs, hedging nouns, and hedging conjunctions. While these lexicons are not exhaustive, the words which have been included are the most frequent hedging words found in many types of literature and their classification as hedgers is supported by linguistic theory and empirical investigations.

Even though hedging devices appear to be used similarly among fraudulent 10Ks and non-fraudulent ones, the warrants linking deception and hedging appear to be justified theoretically. Modest success at classifying fraudulent 10Ks was achieved. Future research should expand the hedging lexicon to include multi-word phrases. For example, “not readily apparent from other sources” appears to be a common phrase within 10Ks with a linguistic purpose of hedging against unforeseeable conditions. Phrases like this have linguistic purpose and may be diagnostic of communicative intent. Future
research can explore the use of hedging in other financial-related documents, such as press releases, letters from the CEO, 10Qs, etc. Knowledge regarding the use of hedging devices can result in diagnostic tools that bolster the success of audit procedures at countering managerial financial fraud.

REFERENCES

detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and 
Negotiation*, 13, 81-106.

predicting deception in CMC. *JMIS* Vol.20(4) pp.139-165.
