INTELLIGENT SYSTEMS IN ACCOUNTING, FINANCE AND MANAGEMENT
Intell. Sys. Acc. Fin. Mgmt. **23,** 215–239 (2016)
Published online 24 May 2016 in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/isaf.1392

DO SENTIMENTS MATTER IN FRAUD DETECTION? ESTIMATING SEMANTIC ORIENTATION OF ANNUAL REPORTS

SUNITA GOEL[a]* AND OZLEM UZUNER[b]

[a] *Siena College, Accounting & Law, Loundonville, NY USA*

[b] *State University of New York at Albany, Department of Information Studies, Albany, NY USA*

SUMMARY

We present a novel approach for analysing the qualitative content of annual reports. Using natural language processing techniques we determine if sentiment expressed in the text matters in fraud detection. We focus on the Management Discussion and Analysis (MD&A) section of annual reports because of the nonfactual content present in this section, unlike other components of the annual reports. We measure the sentiment expressed in the text on the dimensions of polarity, subjectivity, and intensity and investigate in depth whether truthful and fraudulent MD&As differ in terms of sentiment polarity, sentiment subjectivity and sentiment intensity. Our results show that fraudulent MD&As on average contain three times more positive sentiment and four times more negative sentiment compared with truthful MD&As. This suggests that use of both positive and negative sentiment is more pronounced in fraudulent MD&As. We further find that, compared with truthful MD&As, fraudulent MD&As contain a greater proportion of subjective content than objective content. This suggests that the use of subjectivity clues such as presence of too many adjectives and adverbs could be an indicator of fraud. Clear cases of fraud show a higher intensity of sentiment exhibited by more use of adverbs in the "adverb modifying adjective" pattern. Based on the results of this study, frequent use of intensifiers, particularly in this pattern, could be another indicator of fraud. Moreover, the dimensions of subjectivity and intensity help in accurately classifying borderline examples of MD&As (that are equal in sentiment polarity) into fraudulent and truthful categories. When taken together, these findings suggest that fraudulent MD&As in contrast to truthful MD&As contain higher sentiment content.

Keywords: fraud detection; fraud sentiment classification model; textual analysis; natural language processing; Management Discussion and Analysis

1. INTRODUCTION

Academic research on fraud detection suggests that detecting fraud is a complex problem and no one set of predictors will be always successful in fraud detection. This may be partly due to the fact that once the fraud indicators are publicly known, companies can find ways to outsmart them and find other creative ways to conceal fraud. Further research can focus on finding novel fraud indicators and innovative techniques of detecting fraud. Recently, there seems to be an increasing interest in analysing company disclosures using textual analysis, and some of these studies have produced compelling insights into how linguistic structure of text can be exploited to find indicators of fraud (Goel *et al.,* 2010; Humpherys *et al.,* 2011; Goel & Gangolly, 2012; Skillicorn & Purda, 2012; Purda & Skillicorn, 2014). However, this area is still nascent, and a number of areas remain underexplored, particularly the role of sentiment- and emotion-related constructs in fraud detection. With the current study, we aim to

* Correspondence to: Sunita Goel, Accounting & Law, Siena College, 515 Loudon Road, Loudonville, NY 12211, USA. E-mail: sgoel@siena.edu

fill this void and provide evidence to support a new explanation that further advances our understanding of fraud and extends linguistic approaches to fraud detection by examining the role of sentiment-related linguistic nuances such as polarity, subjectivity and intensity.

Mining text for sentiment (also known as affect or emotion) and understanding how it can affect readers' thinking in some way is important because sentiment can convey incorrect information to others when necessary (Ekman & Friesen, 1982). In the current study, by analysing sentiment-laden words in company disclosures, we explore what role sentiment constructs play in fraud detection and how the distribution of sentiment features varies between fraudulent and truthful disclosures. We believe that annual reports and other forms of corporate disclosures are particularly interesting to investigate for fraud indicators for several reasons. First, it is easier to disguise true facts through this mode of communication as it gives a great sense of visual anonymity to the writer, providing an opportunity to misrepresent facts and mask the cues that might reveal fraud. Second, owing to lack of a direct relationship with the users of corporate disclosures, the writer can possibly disassociate from the feelings of guilt and remorse when concealing facts through this mode of communication. Third, with the rise of digitized information and data processing power, it has become possible to investigate for fraud patterns and cues from large text sets, which was not possible earlier.

Specifically, we examine the Management Discussion and Analysis (MD&A) section of annual reports as this section predominantly expresses opinions and attitudes of management in a text. Even though annual reports contain both factual and nonfactual information, these reports are often perceived as conveying factual information to external and internal users. We focused on MD&A because of the nonfactual content present in this section, unlike the other components of annual reports. Although both factual and nonfactual information presented in annual reports is used for decision-making, users of these reports often perceive nonfactual information with the same or greater interest. When examining MD&As, we seek to recognize sentiments that are conveyed by text and how these sentiments differ between truthful and fraudulent MD&As. In spite of burgeoning growth in sentiment analysis, very little empirical evidence exists on the role of sentiment in fraud detection. With the current study, we aim to fill this void and contribute to the emerging research in textual analysis. Recognition of sentiments conveyed by text can also provide useful insight into how management personates itself when it is committing fraud. As the degree of complexity in accounting standards and disclosure requirements continue to grow for financial reporting, we believe our findings may be useful in developing richer tools for (1) predicting the likelihood of fraud in corporate reports and (2) identifying companies that are at high risk of committing fraud. Consequently, our research should be of interest to auditors, fraud examiners, standard setters, regulators, policy-makers and other users including financial analysts, investors and lending institutions.

The remainder of this paper is structured as follows. Section 2 provides a review of relevant literature, concentrating on the studies using textual analysis for fraud detection. Section 3 discusses our sample and the methodology employed in this study. Section 4 describes sentiment feature sets and reports on the experimental results. Finally, Section 5 presents the concluding observations, including the limitations of this study and possible future work directions.

2. LITERATURE REVIEW

The majority of the work on text analysis using natural language processing (NLP) techniques in accounting has focused on extracting factual information. Recently, a few studies in accounting have looked at extracting nonfactual information, such as opinions, sentiments and emotions (Li, 2006; Das & Chen, 2007; Tetlock, 2007; Hájek & Olej, 2013; Hájek *et al.,* 2013). Approaches to the extraction of

DOI: 10.1002/isaf

sentiment from text vary widely from use of lexicon-based approaches to the use of more sophisticated NLP techniques. To date, the research that looks at the interaction of sentiment and fraud has been scanty. In this study, we aim to fill this gap and contribute to the emerging research in textual analysis for fraud detection. Sections 2.1 and 2.2 discuss highlights of studies that analyse textual sentiment in accounting contexts and for fraud detection; Section 2.3 summarizes the main contributions of our work.

2.1. Previous Work on Sentiment Detection in Accounting

The major thrust of sentiment detection studies has been on examining the relationship between investor sentiment extracted from stock investment forums and stock returns. These studies associated bullish messages with positive sentiment and bearish messages with negative sentiment, and message classification was done by matching positive and negative terms from dictionaries with the tokens of the text. For example, Tumarkin and Whitelaw (2001) performed an event study and value-at-risk analysis to examine if the opinions contained in public Internet financial forums help in predicting stock returns and trading volume. In spite of the anecdotal evidence that messages posted to financial forums can be used to manipulate prices, their findings were consistent with market efficiency that it is market information that influences message board activity rather than vice versa. Antweiler and Frank (2004) examined if the level of message posting or the bullishness of the messages posted on Internet stock message boards can predict stock returns, trading volume and stock market volatility. They employed naive Bayes (NB) and support vector machines (SVMs) to carry out all tests and observed that online forum discussions contain financially relevant information, and sentiment can predict trading volume and volatility across stocks but may not be a good predictor of stock movements. In a similar study, Chua *et al.* (2009) constructed an investor sentiment detection engine to find trends and to predict future patterns in the stock market. The messages posted on Internet stock forums based on the Australian market were classified as buy (positive sentiment), hold (neutral sentiment) or sell (negative sentiment). The authors used a variation of the NB classifier, complement NB (CNB) coupled with information gain for feature selection, and achieved an accuracy of 78.72% compared with 57% accuracy from human annotators and 65.63% accuracy in the baseline achieved with a conventional NB classifier.

Li (2006) examined the implications of corporate annual reports' risk sentiment for future earnings and stock returns. He measured the risk sentiment of annual reports by counting the frequency of words related to risk or uncertainty in the 10-K filings. He found that an increase in risk sentiment is associated with lower future earnings. More specifically, firms with a larger increase in risk sentiment had more negative earnings relative to those firms with little increase in risk sentiment in the 12 months after the annual report filing date. Das and Chen (2007) developed a methodology for extracting small-investor sentiment from stock message boards. They primarily used NB to classify investor sentiments. They achieved best results at 62% when they used a simple majority vote of five distinct classifiers, including SVMs. They noted that employment of multiple classifiers to the task of sentiment classification substantially reduced the number of false positives on the one hand and improved the accuracy of the sentiment index on the other hand.

Tetlock (2007) examined the impact of media on future stock prices. He constructed his text sentiment measure from the content of a *Wall Street Journal* column by counting the number of words classified as negative in the Harvard IV-4 dictionary. He used content analysis software General Inquirer together with principal component analysis and found that high pessimism in the media is associated with downward movement of stock market prices. In addition, he noted that unusually high or low pessimism leads to temporarily high market trading volume. His

DOI: 10.1002/isaf

findings were consistent with sentiment theories that media content is linked to the behaviour of investors. Furthermore, he suggests that media content can serve as a proxy for investor sentiment and noninformational trading. In a related study, Garcia (2013) constructed a proxy for market sentiment by counting the number of positive and negative words from two financial columns of the *New York Times* using the Loughran and McDonald (2011) word dictionaries. The author studied its relationship to stock returns and found that investor sentiment has a prominent role during recessionary periods than during expansionary periods. In contrast to Tetlock's study, which focused only on negative word counts, Garcia's study demonstrated that positive words can be as important as negative words and that positive word counts helped in predicting stock returns. Using Loughran and McDonald word dictionaries, Hájek and Olej (2013) evaluated sentiment in annual reports and demonstrated that sentiment information significantly improves the accuracy of financial distress forecasting models. In another study, Hájek *et al.* (2013) combined financial indicators with sentiment indicators obtained from annual reports to improve the accuracy of stock price forecasting models.

2.2. Previous Work on Fraud Detection

Given the complexity of fraud detection and lack of a universal set of fraud indicators, we discuss highlights of studies that have successfully used textual analysis for fraud detection in accounting and other disciplines to gain a richer understanding of language features that have been found to be particularly effective for fraud detection. Burgoon *et al.* (1996) noted that content manipulation is strategic and deceivers intentionally manipulate message information to make it less complete, less clear, less relevant, less direct and less personalized. Burgoon *et al.* (2003) conducted linguistic analysis to distinguish truthful communications from deceptive ones. The results showed that deceivers were more likely to use longer messages but with less diversity and complexity, and greater uncertainty. Newman *et al.* (2003) investigated the features of linguistic style that distinguish between true and false stories. The results of a computer-based text analysis program showed that compared with truth-tellers, liars demonstrated lower cognitive complexity, used fewer self-references and used more negative emotion words. Carlson *et al.* (2004) integrated a broad range of literature on deception, including theories of interpersonal communication and media use, and presented an integrated model of interpersonal deception. They noted that linguistic features can be valid indicators of deception as several of these cues are the result of anxiety, negative emotional states and cognitive demand that occurs in deception. They further pointed out that language choices would likely be impacted by the extent to which communicators are able to plan, craft and edit a message before transmission.

Zhou *et al.* (2004) experimented with automating detection of linguistic cues that are relatively context insensitive and that are often used for deception in text-based asynchronous computer communication where the lie can be carefully prepared beforehand, such as emails, as opposed to synchronous communication where the lie has to be made on the spot, such as Instant Messaging. Nine linguistics constructs – quantity, diversity, complexity, specificity, expressivity, informality, affect, uncertainty and nonimmediacy – were proposed and the results of the experiment showed that a systematic analysis of linguistic information could be useful in differentiating deception from truth. Hancock (2007) investigated changes in linguistic style across truthful and deceptive communication in a synchronous text-based setting. They noted that liars produced more words and used fewer self-oriented pronouns. In a similar study, Hancock *et al.* (2010)

DOI: 10.1002/isaf

examined the role of the communication medium and liar motivation on deception detection and noted that the most difficult liar to catch is one who is highly motivated and communicating in a text-based medium.

2.3. Main Contributions of Present Work

Our work relates closely to financial statement fraud (hereafter, fraud) detection literature that investigates textual content using NLP tools. The use of qualitative predictors for fraud detection has gained much attention in recent years and has provided us with some valuable insights into the importance of linguistic markers in explaining fraud (Cecchini *et al.,* 2010; Goel *et al.,* 2010; Humpherys *et al.,* 2011; Goel & Gangolly, 2012; Skillicorn & Purda, 2012; Goel, 2014; Purda & Skillicorn, 2014). Goel *et al.* (2010), for example, using linguistic features found evidence that there are systematic differences in the communication and writing style of fraudulent and nonfraudulent annual reports. Focusing on the textual content of the MD&A section, Churyk *et al.* (2009), found contextual differences such as use of lexical diversity, present tense verbs and optimism between the MD&A of fraudulent firms and the MD&A of a matched sample of nonfraudulent firms. The authors derived the contextual differences from the deception detection literature. Glancy and Yadav (2011) also examined the MD&A section and provided evidence that writers of fraudulent filings have prior knowledge of fraud. In a related study, Humpherys *et al.* (2011) showed that fraudulent MD&A disclosures in contrast to nonfraudulent disclosures use more activation language, words, imagery, pleasantness, group references and exhibit less lexical diversity.

In this paper we extend this line of qualitative textual research and explore the role of sentiment in fraud detection by investigating sentiments that manifest themselves in the form of linguistic expressions and are evident from the written text. Despite the fact that there is a growing body of work on sentiment analysis, prior empirical work in accounting has not examined the role of sentiment in fraud detection. With the present work we aim to fill this gap and provide another intuitive way to examine the qualitative content of annual reports and determine if sentiment matters in fraud detection. Moreover, our definition of sentiment task is different from how previous studies in accounting and finance have defined sentiment. In particular, we measure sentiment on dimensions of polarity, subjectivity and intensity to examine different ways in which sentiment can be employed to avoid discovery of truth. With the polarity dimension of sentiment, we explore whether fraudulent MD&As are distinguishably more positive or more negative or more neutral than truthful MD&As. Subjectivity refers to aspects of language that are used to express opinions and evaluations (Banfield, 1982; Fludernik, 1993; Wiebe, 1994). With the subjectivity dimension of sentiment, we examine differences in the distribution of subjective sentiment expressions and explore whether fraudulent MD&As are distinguishably more subjective than truthful MD&As. The intensity dimension of sentiment measures the strength of sentiment, and with this we explore whether fraudulent MD&As distinguishably use stronger expressions of sentiments than truthful MD&As do. We believe that fraudulent writings, in contrast to truthful writings, tend to express sentiment differently as the intent of the writer in fraudulent writings is to somehow mask the truth. By analysing sentiment in terms of sentiment polarity, sentiment subjectivity and sentiment intensity we attempt to (1) examine how the distribution of these features varies across fraudulent and truthful MD&As and (2) determine which of these features are strong predictors of fraudulent MD&As.

DOI: 10.1002/isaf

3. RESEARCH METHODOLOGY

3.1. Data, Sample Selection and Preprocessing

In this study, we used Lexis-Nexis, Compustat, *The Wall Street Journal*, *The New York Times*, *The Financial Times* and Accounting and Auditing Enforcement Releases (AAERs) issued by the Securities and Exchange Commission (SEC) to identify companies that were convicted of fraudulent financial reporting during the period 1994 to 2012. Several studies have used AAERs as a proxy for financial statement fraud (Hennes *et al.,* 2008; Cecchini *et al.,* 2010; Dechow *et al.,* 2011). In order to identify AAERs relating to financial statement fraud from the AAER filings listed on the SEC website, we ran keyword searches for phrases such as "Accounting fraud", "fraudulent financial statements and filings", and "Improper and Fraudulent Accounting Entries". When examining these AAERs, we excluded those AAERs that did not include a violation of Rule 17(a) of the Securities Exchange Act of 1933 or a violation of Rule 10(b)-5 of the Securities Exchange Act of 1934. It should be noted that many of the alleged companies had more than one AAER issued to them by the SEC during the course of their investigation, but we included each fraudulent company only once in our sample even if multiple AAERs have been issued to them.

Furthermore, we selected only those fraudulent companies in the final sample that met the fivefold criteria: (1) alleged fraud should affect the annual reports (also called 10-Ks) of the selected companies; (2) the 10-Ks of the selected companies should be available for download from the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) database; (3) original 10-Ks and not the amended 10-Ks of the selected companies should be available for download from EDGAR; (4) the 10-Ks of the selected companies must include an MD&A section; and (5) the selected companies should have a documented evidence of fraud.

This resulted in 180 fraudulent companies identified during the sample period 1994–2012, with 180 fraud-year observations corresponding to the final year of the alleged fraud period of the selected companies. Consistent with previous studies (Cecchini *et al.,* 2010; Humpherys *et al.,* 2011; Goel, 2014), we adopted a matched sample approach in this study. We selected an equal number of nonfraudulent companies from the sample period. Nonfraudulent companies were matched on size (total assets value), industry (Standard Industrial Classification (SIC) code) and time period (year). Thus, for each fraudulent company, a nonfraudulent company was selected from the same industry as defined by the two-digit SIC codes, where available, and the North American Industry Classification System (NAICS) codes otherwise. SIC is a coding system that was developed by the US government for classifying industries. The SIC system is divided into five levels. Letters A–K indicate main divisions. Two-digit SIC codes indicate major groups within the main divisions. Four-digit SIC codes indicate specific industries within the major groups. Six-digit SIC codes indicate sub-industries within the specific industries. And eight-digit SIC codes indicate lines of business within the sub-industries. The NAICS is a new system of coding that was introduced as a replacement for the SIC system. However, certain government departments and agencies, such as the SEC, still use the SIC codes. Further, we verified that the selected nonfraudulent companies have no documented evidence of fraud.

Our final sample consisted of 180 fraudulent and 180 nonfraudulent companies. We next downloaded the 10-Ks of the sample companies. We then extracted the MD&As from the 10-Ks of the selected sample companies. As a result, the final MD&A text corpus comprised 360 MD&As with text that was directly contained in the body of the MD&As. Once the MD&A text corpus was created we performed two preprocessing steps. First, we removed all the tables containing financial information from the text of the MD&As. Second, we removed all the numeric information embedded in the text of

DOI: 10.1002/isaf

the MD&As as numbers do not carry any sentiment information. After preprocessing, the average length of an MD&A text was 12,843 word tokens in the MD&A corpus. We noticed that the length of the MD&A text increased after the SEC issued its guidance on MD&A disclosure in 2003 requiring companies to also discuss critical accounting estimates in the MD&A section.

3.2. Methodology

In this study we employ NLP and machine learning techniques to explore the role of sentiment in fraud detection. In order to identify sentiment features that are useful for differentiating between fraudulent and truthful MD&As, we experiment with sentiment lexicon-based features and more sophisticated linguistic features such as part of speech features. In our experiments, we analyse these features, evaluate their relative predictive capabilities in fraud detection, isolate features that need deeper investigation and subsequently retain only those features that have the most predictive power in distinguishing fraudulent MD&As from truthful MD&As. We use χ^2 and information gain methods for feature ranking in order to identify features that contribute most to the task of fraud detection.

We measure sentiment on dimensions of polarity, subjectivity and intensity (see Section 2.3). Measuring sentiment on these dimensions allowed us to recognize the linguistic choices that the writer makes to influence the sentiment conveyed through text in cases of fraud as two documents may be semantically similar but they may be different in terms of polarity, subjectivity and intensity of sentiments. For example, the words "good", "best" and "outstanding" are semantically similar but they differ in terms of intensity.

We took a lexicon-based approach (also known as dictionary-based approach) to automatically recognize sentiment features for at least three reasons. First, it is not feasible to hand annotate the entire MD&A dataset owing to its large size as this is a labour-intensive and time-consuming process and is more suitable for smaller datasets. Second, automatic recognition of sentiment features as opposed to hand annotation of the MD&A dataset with sentiment labels is less prone to errors and free from personal bias. Third, the availability of lexicon resources, including domain-specific lexicons, has made it possible to automatically recognize sentiment features; for example, see Loughran and McDonald (2011).

In order to extract sentiment features from the MD&A dataset for further analysis we used DICTION 5.0 (Hart, 2000), LIWC (Pennebaker *et al.,* 2007), and Illinois Part-of-Speech tagger (Roth & Zelenko, 1998). DICTION is a textual analysis program that, in addition to standard dictionaries (built-in wordlists), allows for the creation of custom dictionaries and outputs reports containing results about the texts that it processes. The processed output of DICTION includes general language statistics such as total words analysed, total characters analysed, average word size, and the number of different words. It also provides special counts of orthographic characters and high-frequency words. In addition, it reports standard dictionary totals, including raw frequencies, percentages and standardized scores for DICTION's built-in wordlists and provides totals for custom dictionaries. Before processing the MD&A dataset through DICTION, we created custom dictionaries in DICTION for each predefined sentiment lexicon category that was used in this study (see Section 4.1) with the exception of LIWC categories. We used the LIWC tool to extract LIWC category-based sentiment features.

Furthermore, when extracting sentiment features from the MD&A corpus, instead of using binary presence or absence representation of sentiment features, we use their frequency counts. We then normalize these frequency counts before using them in experiments. This is because the length of each MD&A varies in the MD&A corpus and such a variance in length can result in higher frequency counts for longer MD&As relative to shorter MD&As. We discuss lexicon-based features and part-of-speech (POS) features and the process of extracting them in detail in Section 4.

DOI: 10.1002/isaf

We use an SVM, a supervised machine learning method (Joachims, 1998, 2001) to build a fraud sentiment classification model. We selected an SVM as it has been shown that SVMs typically achieve best performance for text categorization tasks out of the other popular classification algorithms, such as NB and maximum entropy (Bradley *et al.,* 1998; Joachims, 1998; Dumais *et al.,* 1998; Yu *et al.,* 2003). We use the Waikato Environment for Knowledge Analysis (WEKA) platform to build the fraud sentiment classification model (Witten & Frank, 2005). Moreover, for the SVM implementation in WEKA we use the sequential minimal optimization method (Platt, 1998) with a linear kernel. We train the SVM with sentiment features on the labelled MD&A dataset, which contains equal numbers of positive (180 fraudulent MD&As) and negative (180 truthful MD&As) examples. The text of each MD&A was converted into a feature vector consisting of sentiment features occurring in that MD&A. Moreover, the sentiment features were normalized before being passed to the input vector of the fraud sentiment classification model. The fraud sentiment classification experiments were performed using a 10-fold cross-validation (CV) technique. Tenfold CV first divides the dataset into 10 folds, and then uses nine folds for training the model and one fold for testing. This process is repeated 10 times, with a different fold being used each time for evaluating the model performance and the remaining nine folds used for training.

The accuracy of the fraud sentiment classification model was calculated in terms of correctly classified instances; that is, true positives and true negatives divided by the total number of classified instances (i.e. true positives, false positives, true negatives and false negatives). We define true positives as the number of fraudulent MD&As that are accurately classified as fraudulent by the model and define false positives as the number of truthful MD&As that are inaccurately classified as fraudulent by the model. On the other hand, we define true negatives as the number of truthful MD&As that are accurately classified as truthful by the model and define false negatives as the number of fraudulent MD&As that are inaccurately classified as truthful by the model. In this study, we report the average results of the classification performance across the 10 folds for all the experiments.

In addition, for all the experiments, we report the true-positive rate

$$\frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

the false-positive rate

$$\frac{\text{false positives}}{\text{false positives} + \text{true negatives}}$$

precision, recall and *F*-score results for each class (fraudulent and truthful) separately. Recall measures the proportion of true cases that were correctly identified, and precision measures the proportion of the predicted cases that were correct. The *F*-score

$$2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

shows the balance between precision and recall. Performance measures such as recall, precision, and *F*-score in addition to performance accuracy help us assess if the classification model built in this study is good enough to solve our domain problem. Accuracy by itself may not be sufficient, as a

DOI: 10.1002/isaf

model that is selected based on high accuracy can still have low predictive power on the domain problem and, as such, should not be selected. We next discuss the feature sets and the classification experiments with them.

4. FEATURE SETS AND EXPERIMENTS

4.1. Sentiment Lexicon-based Feature Set

Sentiment lexicons are referred to as lexical resources that contain words and their associated semantic orientation. By matching input words in the text against predefined lists of words annotated with sentiment labels in sentiment lexicons, it is possible to determine the semantic orientation of input words in the text. After examining several sentiment lexicons, dictionaries and sentiment analysis tools, we selected only those lexicons and tools for automatic sentiment feature recognition that were either specific to our domain or were found to be robust across multiple domains. In particular, we use the Loughran and McDonald (2011) financial sentiment dictionary as it is specific to our domain. In addition, we use two other lexicons, namely the Multi-Perspective Question Answering (MPQA) subjectivity lexicon (Wilson *et al.,* 2005) and the Linguistic Inquiry and Word Count (LIWC; Pennebaker *et al.,* 2007), as the performance of these two lexicons has been shown to be robust across multiple domains (Devitt & Ahmad, 2007; Ding *et al.,* 2008; Taboada *et al.,* 2011). We did not select other popular lexicons, such as SentiWordNet (Esuli & Sebastiani, 2006), Whissell's dictionary of affect in language (Whissell, 1989), General Inquirer (Stone *et al.,* 1966) or the SentiStrength lexicon (Thelwall *et al.,* 2012), as these were found not to be suitable for our task. For example, lots of words used in the original General Inquirer positive and negative wordlists have different connotations in the financial domain. Similarly, SentiStrength, for example, was not selected as it was found to be ideal for short texts and in order to use it for our task it would require additional preprocessing where the text of each MD&A document is first broken down into smaller size chunks and then the results are aggregated at the document level for each MD&A. We next discuss lexicon features and the approach that we took to extract them using each of the three selected lexicons.

4.1.1. Loughran and McDonald Financial Sentiment Dictionary

We use the Loughran and McDonald (hereafter, LM) financial sentiment dictionary as it is specific to our domain of interest and includes commonly occurring words in financial text. Moreover, the use of domain-specific dictionaries improves the quality of analysis and leads to fewer misclassifications when determining the correct semantic orientation of terms in a financial context (Loughran & McDonald, 2011; O'Leary, 2011). The LM financial sentiment dictionary contains six categorical predetermined wordlists that are suggested to capture sentiment: *positive*, *negative*, *uncertainty*, *litigious*, *modal strong* and *modal weak*. We created custom dictionaries based on the LM categorization for *positive*, *negative*, *modal strong* and *modal weak* wordlists in DICTION. Close examination of the other two LM wordlists namely *uncertainty* and *litigious* revealed that these wordlists are not consistent with the conceptual definition of the features that express sentiment polarity, sentiment subjectivity, and sentiment intensity and as a result, were not used in this study. Examples of *LM positive* words include "excellent", "innovative"; *LM negative* words include "bankruptcy", "discontinued", "felony", "loss", "misstated"; *LM uncertainty* words include "risk", "assume", "ambiguity"; *LM litigious* words include "defendant", "plaintiff", "testimony", "breach"; *LM modal strong* words include words "best", "highest", "lowest"; and *LM modal weak* words include "possibly" and "sometimes".

DOI: 10.1002/isaf

We then processed fraudulent and truthful MD&A corpora through DICTION to extract sentiment features for each of these LM categories. We refer to these features as LM features (see Table I). To do this, we activated custom dictionaries *LM positive*, *LM negative*, *LM modal strong*, and *LM modal weak* such that they were invoked each time DICTION software analysis was run to process the MD&A text corpus consisting of 180 fraudulent and 180 truthful MD&As. In other words, for each MD&A in the corpus, a vector of *LM positive*, *LM negative*, *LM modal strong* and *LM modal weak* category counts was derived from DICTION. We then normalized these individual LM category counts by the length of the MD&A before running classification experiments with them.

4.1.2. Multi-Perspective Question Answering Subjectivity Lexicon

We use the MPQA subjectivity lexicon, which is a part of Opinion Finder system (Wilson *et al.,* 2005) to primarily measure subjective sentiment. Out of all the lexicons that we examined, this is the only resource that annotates words with subjectivity markers. All the entries in this lexicon have been labelled as either strong (strongsubj) or weak (weaksubj) indicators of subjectivity depending upon the frequency with which each appears in subjectivity contexts. Specifically, words that are subjective in most contexts are marked as strongly subjective, whereas words that only have certain subjective usages are marked as weakly subjective (Wilson *et al.,* 2005). In addition, for each subjective word in the lexicon, information about its prior polarity, POS and stemming is also included (see Figure 1) and each word in the respective polarity is labelled as positive, negative or neutral. On a closer examination we noticed that several words labelled with neutral polarity were in fact strong indicators of subjectivity because they indicated a different type of subjective expression, such as speculation.

In order to capture subjective content using the MPQA lexicon and also to collect information on polarity of subjective language identified using MPQA, we extracted words tagged with each type of polarity using a shell script and grouped them according to their polarity into positive, negative and neutral categories. Since all the words included in the MPQA lexicon are subjective words with varying degree of subjectivity, we also extracted all subjective words contained in the MPQA Lexicon using a shell script and grouped them under the subjective category. For each of these extracted categories, we then created custom dictionaries called *MPQA subjective*, *MPQA positive*, *MPQA negative* and *MPQA neutral* in DICTION. Some subjective categories in the MPQA lexicon had multiple instances of the same subjective word corresponding to different parts of speech, but we only included it once in the relevant custom dictionaries (see Figure 1). We then processed fraudulent and truthful MD&A corpora through DICTION

Table I. Summary of LM features used to build fraud sentiment classification model

Sentiment feature	Sentiment dimension	Description
LM_Positive	Polarity dimension of sentiment	Frequency count of occurrence of words on the positive wordlist of LM Financial Sentiment Dictionary scaled by the total number of words in the MD&A text document being analysed
LM_Negative	Polarity dimension of sentiment	Frequency count of occurrence of words on the negative wordlist of LM Financial Sentiment Dictionary scaled by the total number of words in the MD&A text document being analysed
LM_ Modal_Strong	Intensity dimension of sentiment	Frequency count of occurrence of words on the "modal strong" wordlist of LM Financial Sentiment Dictionary scaled by the total number of words in the MD&A text document being analysed
LM_Modal_Weak	Intensity dimension of sentiment	Frequency count of occurrence of words on the "modal weak" wordlist of LM Financial Sentiment Dictionary scaled by the total number of words in the MD&A text document being analysed

DOI: 10.1002/isaf

(a) Snippet from MPQA Subjectivity Lexicon

```
type=strongsubj len=1 word1=absolute pos1=adj stemmed1=n
priorpolarity=neutral
type=strongsubj len=1 word1=absolutely pos1=adj stemmed1=n
priorpolarity=neutral
type=weaksubj len=1 word1=affect pos1=verb stemmed1=y
priorpolarity=neutral
type=strongsubj len=1 word1=affected pos1=adj stemmed1=n
priorpolarity=neutral
type=weaksubj len=1 word1=risk pos1=adj stemmed1=n
priorpolarity=negative
type=weaksubj len=1 word1=risk pos1=noun stemmed1=n
priorpolarity=negative
```

(b) Custom Dictionaries for the MPQA Subjectivity Lexicon Snippet shown in (a)

MPQA Subjective	MPQA Positive	MPQA Negative	MPQA Neutral
absolute		risk	absolute
absolutely			absolutely
affect			affect
affected			affected
risk			

Figure 1. Illustration of creation of MPQA custom dictionaries using MPQA subjectivity lexicon: (a) snippet from MPQA subjectivity lexicon; (b) custom dictionaries for the MPQA subjectivity lexicon snippet shown in (a).

to extract sentiment features for each of these MPQA-based custom dictionaries. We refer to these features as MPQA features (see Table II). Thus, for each MD&A in the corpus, a vector of category counts – *MPQA subjective* words count, *MPQA positive* words count, *MPQA negative* words count, and *MPQA neutral* words count – was obtained from DICTION. We then normalized these individual MPQA category counts by the length of the MD&A before running classification experiments with them.

4.1.3. Linguistic Inquiry and Word Count Categories

In this study, we also use LIWC, a computer-aided text analysis program to extract sentiment features. Use of LIWC is motivated by the fact that it examines emotional, cognitive and structural components of text in detail by organizing the occurrence of words into different lexical and content categories, and several of these categories reflect the emotional and mental states of the writer (Pennebaker *et al.,* 2007). In particular, LIWC analyses each incoming word by matching it with words in its 82 categories,

Table II. Summary of MPQA features used to build fraud sentiment classification model

Sentiment feature	Sentiment dimension	Description
MPQA_Subjective	Subjectivity dimension of sentiment	Frequency count of occurrence of words on the MPQA subjective custom wordlist scaled by the total number of words in the MD&A text document being analysed
MPQA_Positive	Polarity dimension of sentiment	Frequency count of occurrence of words on the MPQA positive custom wordlist scaled by the total number of words in the MD&A text document being analysed
MPQA_Negative	Polarity dimension of sentiment	Frequency count of occurrence of words on the MPQA negative custom wordlist scaled by the total number of words in the MD&A text document being analysed
MPQA_Neutral	Polarity dimension of sentiment	Frequency count of occurrence of words on the MPQA neutral custom wordlist scaled by the total number of words in the MD&A text document being analysed

DOI: 10.1002/isaf

and if a match is found the scores of corresponding categories are incremented. In this study, in order to examine if there is a tendency to use sentiment dimension of language differently by the writers of the fraudulent and truthful MD&As, we selected a subset of LIWC categories and subcategories; namely, *Posemo* (positive emotion subcategory of Affect category), *Negemo* (negative emotion subcategory of Affect category), *Anx* (anxiety subcategory of Affect category), *Anger* (subcategory of Affect category), *Sad* (subcategory of Affect category), *Affect* (category). We then processed fraudulent and truthful MD&A corpora through the LIWC tool to extract features for each of these six categories. We refer to these features as LIWC features (see Table III). Once the entire MD&A corpus was processed, the output of LIWC provided us with relative frequencies of LIWC features for each MD&A.We also present the distribution of MD&A corpora as per LIWC positive and negative emotion scores (see Table IV). Close examination of the distribution of LIWC features indicated that some of them, such as anger, anxiety and sadness, are very infrequent in the MD&A corpus. Even though these features have been suggested to be emotion-bearing features in previous studies, the nonzero values of these features occur only 3% of the time in the MD&A corpus.

Table III. Summary of LIWC features used to build fraud sentiment classification model

Sentiment feature	Sentiment dimension	Description
LIWC_Positive_Emotion	Polarity dimension of sentiment	Relative frequency of the "Posemo" category words expressed as a percentage of LIWC total words of the MD&A text document being analysed
LIWC_Negative_Emotion	Polarity dimension of sentiment	Relative frequency of the "Negemo" category words expressed as a percentage of LIWC total words of the MD&A text document being analysed
LIWC_Anxiety_Emotion	Polarity dimension of sentiment	Relative frequency of the "Anx" category words expressed as a percentage of LIWC total words of the MD&A text document being analysed
LIWC_Anger_Emotion	Polarity dimension of sentiment	Relative frequency of the "Anger" category words expressed as a percentage of LIWC total words of the MD&A text document being analysed
LIWC_Sadness_Emotion	Polarity dimension of sentiment	Relative frequency of the "Sad" category words expressed as a percentage of LIWC total words of the MD&A text document being analysed
LIWC_Affect	Polarity dimension of sentiment	Relative frequency of the "Affect" category words expressed as a percentage of LIWC total words of the MD&A text document being analysed

Table IV. Distribution of MD&A corpora based on LIWC positive and negative emotion scores

MD&A categorization based on LIWC emotion scores	Fraudulent MD&As (180)	Truthful MD&As (180)
Count of MD&As that do not possess any positive or negative emotion	2	7
Count of MD&As that possess almost equal positive and negative emotions	37	53
Count of MD&As that possess higher positive emotion score as opposed to negative emotion score	88	74
Count of MD&As that possess higher negative emotion score as opposed to positive emotion score	53	46

DOI: 10.1002/isaf

4.2. Part-of-Speech-based Feature Set

In addition to lexicon-based sentiment features, we also examine the role of more sophisticated linguistic features, such as POS (namely nouns, verbs, adjectives, adverbs, superlative adjectives, comparative adjectives, superlative adverbs, comparative adverbs, and different combinations of these), to identify sentiment features that can be used to distinguish fraudulent MD&As from truthful MD&As. The use of POS features in this paper was motivated by three reasons. First, since use of lexicon-based sentiment features was limited to the intersection of dictionary entries and MD&A text, use of POS features allowed us to examine occurrences of out-of-dictionary words; that is, words that may not be present in any of the pre-established sentiment dictionaries or the sentiment lexicon that was used to extract lexicon-based sentiment features. Second, an overwhelming body of work provides empirical evidence that POS information is useful in identifying markers of sentiment and in determining semantic orientation of text (Bruce & Wiebe, 1999; Wiebe *et al.,* 1999; Benamara *et al.,* 2007; Taboada *et al.,* 2011). Third, since the MD&A section of an annual report contains a higher proportion of nonfactual content compared with the other sections of the annual report, we expect to find a higher percentage of expressions of sentiment in this section. However, some of these expressions of sentiment may not necessarily contain any sentiment words. Sometimes writers use emphasis words to make assertive statements that focus a user's attention on certain topics. Therefore, emphasis words have been shown to express higher emotional involvement with content (e.g. Liao *et al.,* 2008; Guo, 2014). These emphasis words are often captured by adjectives, adverbs, nouns and verbs that express subjective language and opinions. With the help of POS information, we can quantify such expressions of sentiment in text. For example, in

> While we may continue to make selected complementary acquisitions, we *anticipate* that the amount of acquisition activity will be *significantly* reduced, and, *therefore*, *expect* that our growth rate in revenues and earnings from acquisitions will *also* be reduced as compared to prior periods.

a sample sentence extracted from the MD&A corpus, the intensity word "significantly" is an adverb, as are the subjective opinion word "therefore" and the emphasis word "also". Furthermore, the subjective words "anticipate" and "expect" in the sample MD&A sentence are verbs.

As a result, in this study, we explore POS categories such as nouns, verbs, adjectives, adverbs, superlative adjectives, comparative adjectives, superlative adverbs, comparative adverbs and different combinations of these POS features to identify sentiment markers in MD&A corpus. Nouns represent names of entities, whereas verbs characterize actions or states. Adjectives describe a property or attribute of the entities and typically have a base form, an adverb form, a comparative form and a superlative form. Similarly, adverbs describe a property or an attribute of the actions or states denoted by verbs and typically have a base form, a comparative form and a superlative form.

Using a POS tagger (Roth & Zelenko, 1998) we obtained the POS annotation of each word for the entire MD&A corpus (see Figure 2a for an example of an input to the POS tagger and see Figure 2b for the corresponding output). This tagger inserts the POS tag (see Figure 2c for a list of POS tags used by the tagger to annotate the input example shown in Figure 2a) for each word right in front of it; for example, "NN Accounting" indicates that "Accounting" is a singular noun (marked by "NN"). Once the entire MD&A corpus was annotated with POS tags, for each MD&A we derived POS features for the following nine POS categories (as shown in Figure 2d):

1 *Nouns*. This category contained combined counts of singular nouns and plural nouns. Based on preliminary observations, we noted that counting singular and plural nouns separately does not carry any useful information for our task of sentiment analysis in terms of polarity, subjectivity and intensity.

DOI: 10.1002/isaf

(a) Excerpt from a partial MD&A

Accounting estimates and assumptions discussed in this section are those that we consider to be the most critical to an understanding of our financial statements because they inherently involve significant judgments and uncertainties. All of these estimates reflect our best judgment about current, and for some estimates future, economic and market conditions and their effects based on information available as of the date of these financial statements. If such conditions persist longer or deteriorate further than expected, it is reasonably possible that the judgments and estimates described below could change, which may result in future impairments of investment securities, goodwill, intangibles and long-lived assets, incremental losses on financing receivables, establishment of valuation allowances on deferred tax assets and increased tax liabilities, among other effects.

(b) POS Tagger Output for the MD&A excerpt shown in (a)

NN Accounting NNS estimates CC and NNS assumptions VBN discussed IN in DT this NN section VBP are DT those IN that PRP we VBP consider TO to VB be DT the RBS most JJ critical TO to DT an NN understanding IN of PRP$ our JJ financial NNS statements IN because PRP they RB inherently VBP involve JJ significant NNS judgments CC and NNS uncertainties . . DT All IN of DT these NNS estimates VBP reflect PRP$ our JJS best NN judgment IN about JJ current , , CC and IN for DT some NNS estimates JJ future , , JJ economic CC and NN market NNS conditions CC and PRP$ their NNS effects VBN based IN on NN information JJ available IN as IN of DT the NN date IN of DT these JJ financial NNS statements . . IN If JJ such NNS conditions VBP persist RBR longer CC or VB deteriorate JJ further IN than VBN expected , , PRP it VBZ is RB reasonably JJ possible IN that DT the NNS judgments CC and NNS estimates VBN described RB below MD could VB change , , WDT which MD may VB result IN in JJ future NNS impairments IN of NN investment NNS securities , , NN goodwil l , , VBZ intangibles CC and JJ long-
lived NNS assets , , JJ incremental NNS losses IN on NN financing NN receivables , , NN establishment IN of NN valuation NNS allowances IN on JJ deferred NN tax NNS assets CC and VBN increased NN tax NNS liabilitie s , , IN among JJ other NNS effects.

(c) Key for the sample POS Tagger Output in (b)

DT	Determiner	NNS	Plural noun	RB	Adverb
IN	Preposition	JJ	Adjective	TO	to
NN	Singular noun	VBP	Verb, non 3rd ps. sing. Present	MD	Modal
VBN	Verb, past participle	CC	Coordinating conjunction	WDT	wh-determiner
PRP	Personal pronoun`	VBZ	Verb, 3rd ps. sing. present	RBS	Superlative adverb
PRP$	Possessive pronoun	JJS	Superlative adjective	VB	Verb, base form
RBR	Comparative adverb				

(d) Frequency counts of POS features for the sample POS Tagger Output in (b)

	Count		Count
Nouns (including NN, NNS)	36	Adverbs (RB)	3
Verbs (including VB, VBP, VBN, VBZ)	16	Comparative Adverbs (RBR)	1
Adjectives (JJ)	16	Superlative Adverbs (RBS)	1
Superlative Adjectives (JJS)	1	Comparative Adjectives (JJR)	0
Adverbs modifying Adjectives (RB, JJ)	1		

Figure 2. Illustration of extraction of POS features for a partial MD&A document: (a) excerpt from a partial MD&A; (b) POS tagger output for the MD&A excerpt shown in (a); (c) key for the sample POS tagger output in (b); (d) frequency counts of POS features for the sample POS tagger output in (b).

2 *Verbs*. This contained combined counts of base form of verb, past tense form of verb, past participle form of verb, gerund/present participle form of verb, third-person singular present form of verb, non-third-person singular present form of verb. We did not count different tenses of verbs separately as we do not need this level of granularity for our task of sentiment analysis.
3 *Adjectives*, which we expect to mark subjectivity.
4 *Comparative adjectives*, which we expect to mark lack of subjectivity.
5 *Superlative adjectives*, which we expect to mark intensity.
6 *Adverbs*, which we expect to mark subjectivity.
7 *Comparative adverbs*, which we expect to mark intensity.

DOI: 10.1002/isaf

8 *Superlative adverbs*, which we expect to mark intensity.
9 *Adverbs modifying adjectives*. We extracted consecutive words from the POS annotated MD&A corpus matching adverb adjective pattern as this pattern was found to be particularly useful for distinguishing borderline cases of MD&As that were equal in terms of polarity.

We collected information on inflectional forms of adjectives and adverbs separately as comparative and superlative forms helped us in recognizing different sentiment dimensions (subjectivity and intensity). Figure 2 illustrates the process that we use to derive POS features for a partial MD&A document. We then normalize the frequency counts of the POS features by the length of the MD&A before running classification experiments with them. Table V summarizes POS features used in this study.

4.3. Experiments

Using a labelled dataset of 180 fraudulent and 180 truthful MD&As, we construct a fraud sentiment classification model to identify which combination or selection of sentiment features would obtain the highest accuracy in fraud detection. We run additional tests to identify sentiment features that occur with greater frequency in fraudulent MD&As and are strong predictors of fraudulent MD&As. To build the fraud sentiment classification model, we employ a supervised machine learning method, SVM, and train the SVM model with lexicon-based sentiment features and more sophisticated linguistic-based sentiment features extracted using POS tag information. Specifically, each document in the MD&A corpus is represented as a vector of counts of the number of occurrences of each sentiment feature occurring in that document before being passed to the model. The model is then trained on the sentiment feature values (normalized values) of the MD&A documents for fraud classification. The output of the model is the result of the classification; that is, whether a document belongs to a fraudulent class or a truthful class. When examining an unseen document, the model assigns the class whose feature frequency distribution matches best the feature frequency distribution of the unseen document.

Our baseline experiments consisted of only lexicon-based features. For constructing the baseline model we took an incremental approach where we progressively added one sentiment dimension related feature at a time. We started our experiments with sentiment polarity features. We then added sentiment

Table V. Summary of POS features used to build fraud sentiment classification model

Sentiment feature	Sentiment dimension	Description
Nouns	Subjectivity dimension of sentiment	Frequency of nouns scaled by the total number of words in the MD&A text document being analysed
Verbs	Subjectivity dimension of sentiment	Frequency of verbs scaled by the total number of words in the MD&A text document being analysed
Adjectives	Subjectivity dimension of sentiment	Frequency of adjectives scaled by the total number of words in the MD&A text document being analysed
Comparative Adjectives	Subjectivity dimension of sentiment	Frequency of comparative adjectives scaled by the total number of words in the MD&A text document being analysed
Superlative Adjectives	Intensity dimension of sentiment	Frequency of superlative adjectives scaled by the total number of words in the MD&A text document being analysed
Adverbs	Subjectivity dimension of sentiment	Frequency of adverbs scaled by the total number of words in the MD&A text document being analysed
Comparative Adverbs	Intensity dimension of sentiment	Frequency of comparative adverbs scaled by the total number of words in the MD&A text document being analysed
Superlative Adverbs	Intensity dimension of sentiment	Frequency of superlative adverbs scaled by the total number of words in the MD&A text document being analysed
Adverbs modifying Adjectives	Intensity dimension of sentiment	Frequency of adverbs modifying adjectives scaled by the total number of words in the MD&A text document being analysed

DOI: 10.1002/isaf

subjectivity features followed by sentiment intensity features. We report our results in Table VI separately for each of these experiments. Our next set of experiments consisted of POS features only. Encouraged by the results obtained individually with lexicon-based features and with POS features, for our next experiment we combined POS features and lexicon-based features in an attempt to see whether this combined set of features boosts the performance of the model. For all these experiments, we report average 10-fold CV classification accuracy results in Table VI (see Section 3.2 for a discussion of CV), followed by confusion matrix and detailed accuracy results in Tables VII and VIII respectively (see Section 4.3.2). We next discuss these experimental results in detail.

4.3.1. Discussion of Classification Experiments

The first experiment contained lexicon-based sentiment polarity features only, and with these features the model achieved a modest performance of 63.57% (see Table VI). For the second experiment we excluded three LIWC features related to anger, sadness and anxiety from the feature space as we observed that these features were sparsely distributed across the fraudulent and truthful corpora and we wanted to examine whether their exclusion has any effect on the performance of the model. When these LIWC features were withheld, the performance of the model improved from 63.57% to 65.75%, suggesting that their inclusion is not very helpful in the learning process and might in part be causing degrading model performance. Consequently, these three LIWC features were withheld from further experiments.

For the third experiment, subjectivity features were added to the lexicon-based sentiment polarity features. With the inclusion of subjectivity features, the model outperforms the accuracy obtained with polarity features alone and achieves an accuracy of 68.58%. This shows that subjective features besides

Table VI. Fraud sentiment classification accuracy with different sentiment features

Classification model	Sentiment features	Average classification accuracy results of 10-fold CV (%)
Model 1	Lexicon-based polarity features (including all six LIWC features) (LM_Positive, LM_Negative, MPQA_Positive, MPQA_Negative, MPQA_Neutral, LIWC_Positive_Emotion, LIWC_Negative_Emotion, LIWC_Anxiety_Emotion, LIWC_Anger_Emotion, LIWC_Sadness_Emotion, LIWC_Affect)	63.57
Model 2	Lexicon-based polarity features (with only three LIWC features) (LM_Positive, LM_Negative, MPQA_Positive, MPQA_Negative, MPQA_Neutral, LIWC_Positive_Emotion, LIWC_Negative_Emotion, LIWC_Affect)	65.75
Model 3	Lexicon-based polarity features (with only three LIWC features) plus lexicon-based subjectivity features (LM_Positive, LM_Negative, MPQA_Positive, MPQA_Negative, MPQA_Neutral, LIWC_Positive_Emotion, LIWC_Negative_Emotion, LIWC_Affect, MPQA_Subjective)	68.58
Model 4	Lexicon-based polarity features (with only three LIWC features) plus lexicon-based subjectivity features plus lexicon-based intensity features (LM_Positive, LM_Negative, MPQA_Positive, MPQA_Negative, MPQA_Neutral, LIWC_Positive_Emotion, LIWC_Negative_Emotion, LIWC_Affect, MPQA_Subjective, LM_Modal_Strong, LM_Modal_Weak)	71.69
Model 5	POS features (nouns, verbs, adjectives, adverbs, superlative adjectives, comparative adjectives, superlative adverbs, comparative adverbs, adverbs modifying adjectives)	73.98
Model 6	Lexicon-based polarity features (with only three LIWC features) plus lexicon-based subjectivity features plus lexicon-based Intensity features plus POS features (LM_Positive, LM_Negative, MPQA_Positive, MPQA_Negative, MPQA_Neutral, LIWC_Positive_Emotion, LIWC_Negative_Emotion, LIWC_Affect, MPQA_Subjective, LM_Modal_Strong, LM_Modal_Weak, nouns, verbs, adjectives, adverbs, superlative adjectives, comparative adjectives, superlative adverbs, comparative adverbs, adverbs modifying adjectives)	80.15
Model 7	Top 10 features based on χ^2 test (see Table X)	81.84

DOI: 10.1002/isaf

Table VII. Confusion matrix results of fraud sentiment classification models

Class	Fraudulent MD&A (predicted)	Truthful MD&A (predicted)	Total (actual)
Classification model 1 with average accuracy of 63.57%			
Fraudulent MD&A	97	83	180
Truthful MD&A	48	132	180
Total (predicted)	145	215	360
Classification model 2 with average accuracy of 65.75%			
Fraudulent MD&A	102	78	180
Truthful MD&A	45	135	180
Total (predicted)	147	213	360
Classification model 3 with average accuracy of 68.58%			
Fraudulent MD&A	104	76	180
Truthful MD&A	37	143	180
Total (predicted)	141	219	360
Classification model 4 with average accuracy of 71.69%			
Fraudulent MD&A	117	63	180
Truthful MD&A	39	141	180
Total (predicted)	156	204	360
Classification model 5 with average accuracy of 73.98%			
Fraudulent MD&A	124	56	180
Truthful MD&A	38	142	180
Total (predicted)	162	198	360
Classification model 6 with average accuracy of 80.15%			
Fraudulent MD&A	152	28	180
Truthful MD&A	43	137	180
Total (predicted)	195	165	360
Classification model 7 with average accuracy of 81.84%			
Fraudulent MD&A	157	23	180
Truthful MD&A	42	138	180
Total (predicted)	199	161	360

polarity features are also useful. For our unreported experiments with lexicon-based subjective features alone, the classification performance was only 57.55%. This low performance may be because the model may not have enough features to learn for the classification process, which may be responsible for underfitting of the data. Another possible reason may be that lexicon-based subjectivity features by themselves may not be discriminative in fraud detection but when combined with other features can be effective in fraud detection. The fourth experiment contained all lexicon-based features, as we added sentiment intensity features to the existing set of sentiment polarity and sentiment subjectivity features. We noted that the baseline model achieved an accuracy of 71.69% and its performance improved by a narrow margin when intensity features were added. This may be because the distribution of lexicon-based intensity features was more or less uniform across the fraudulent and truthful MD&A corpora.

For a richer exploration of sentiment that goes beyond the lexicon- or dictionary-based features, the fifth experiment contained POS features only. This experiment was also guided by our intuition to

DOI: 10.1002/isaf

Table VIII. Detailed results of fraud sentiment classification models

Class	TP rate	FP rate	Precision	Recall	*F*-score
Classification model 1 with average accuracy of 63.57%					
Fraudulent MD&A	0.5389	0.2667	0.6690	0.5389	0.5969
Truthful MD&A	0.7333	0.4611	0.6140	0.7333	0.6684
Classification model 2 with average accuracy of 65.75%					
Fraudulent MD&A	0.5667	0.2500	0.6939	0.5667	0.6239
Truthful MD&A	0.7500	0.4333	0.6338	0.7500	0.6870
Classification model 3 with average accuracy of 68.58%					
Fraudulent MD&A	0.5778	0.2056	0.7376	0.5778	0.6480
Truthful MD&A	0.7944	0.4222	0.6530	0.7944	0.7168
Classification model 4 with average accuracy of 71.69%					
Fraudulent MD&A	0.6500	0.2167	0.7500	0.6500	0.6964
Truthful MD&A	0.7833	0.3500	0.6912	0.7833	0.7344
Classification model 5 with average accuracy of 73.98%					
Fraudulent MD&A	0.6889	0.2111	0.7654	0.6889	0.7251
Truthful MD&A	0.7889	0.3111	0.7172	0.7889	0.7513
Classification model 6 with average accuracy of 80.15%					
Fraudulent MD&A	0.8444	0.2389	0.7795	0.8444	0.8107
Truthful MD&A	0.7611	0.1556	0.8303	0.7611	0.7942
Classification model 7 with average accuracy of 81.84%					
Fraudulent MD&A	0.8722	0.2333	0.7889	0.8722	0.8285
Truthful MD&A	0.7667	0.1278	0.8571	0.7667	0.8094

TP: true positive; FP: false positive.

compare the baseline model's performance achieved with lexicon-based features with its performance with POS features. The results showed that the model was able to achieve a higher accuracy with POS features alone (73.98%) in contrast to the lexicon-based sentiment features (71.69%).

Inspired by the accuracy achieved with POS features alone, in the sixth experiment we investigated the impact on model performance of combining POS sentiment features with lexicon-based sentiment features. We noted that we were able to improve the predictive accuracy of the model by at least 6% beyond the accuracy achieved individually with POS features (73.98%) or lexicon-based features (71.69%), and the model achieved an accuracy of 80.15% when we combined POS sentiment features with lexicon-based sentiment features. Our seventh experiment was guided by the inclusion of only the top 10 ranking sentiment features as per the χ^2 test (see Table X). The results demonstrate that the model performs the best among all the experiments with these top 10 ranking features selected from the feature space of 23 features and achieves an accuracy of 81.15%. In other words, 81% of the MD&A documents can be correctly classified as either fraudulent or truthful using the model developed in this study. When compared with random baseline classification accuracy of 50% for a two-class problem, the model built in this study is able to accurately identify fraudulent and truthful MD&As 30% better than chance.

DOI: 10.1002/isaf

4.3.2. Confusion Matrix and Detailed Accuracy Results of Classification

We show the confusion matrix and detailed accuracy results for each of the seven classification models listed in Table VI in Tables VII and VIII respectively. The confusion matrix shows the four metrics true positives, false positives, true negatives and false negatives. As seen in Table VII, for classification model 1 the confusion matrix for 10-fold CV shows that the SVM correctly classified 229 instances out of the total of 360 instances (63.57%) and incorrectly classified 131 instances (36.43%). The diagonal cells in the confusion matrix show the number of MD&As that are correctly assigned by the model to their true class. For example, for the first model, the first row in the confusion matrix indicates that there are 180 total instances that should be classified as belonging to the fraud class. The model classified 97 of them correctly as fraudulent MD&As and incorrectly classified 83 fraudulent MD&As as truthful MD&As. The second row indicates that there are 180 total instances that should be classified as belonging to the truthful class. In this case the model classified 132 MD&As correctly and incorrectly classified 48 truthful MD&As as fraudulent MD&As.

Table VIII shows detailed accuracy results for each of the seven classification models in terms of true-positive rate, false-positive rate, precision, recall and *F*-score (see Section 3.2 for explanation of these performance measures). It can be observed that as the classification accuracy improved from 63.57% (achieved with model 1) to 81.84% (achieved with model 7), the predictive power of the model also improved for the domain problem. For instance, it can be seen that the true-positive rate or recall increased for the fraudulent class from 53.89% (model 1) to 87.22% (model 7). More importantly, the false-positive rate for the truthful class went down from 0.4611 (model 1) to 0.1278 (model 7), which means that model 7 only missed 12.78% of the fraudulent cases and misclassified them as truthful whereas model 1 missed 46.11% of the fraudulent cases and misclassified them as truthful. Thus, for our domain problem, the performance of model 7 is far superior when compared with the performance of the other models (1–6). Next, we test if the observed differences between the performance accuracies of the models are statistically significant.

4.3.3. Significance of Differences in Accuracies of Classification Models

We also examine if the observed differences between the performance accuracies of the SVM classification models obtained with different sets of sentiment features are indeed significant. Even though there is no statistical significance test that would satisfy all the constraints, significance tests nevertheless do provide approximate confidence levels, which can help in interpreting experimental comparisons of models (Mitchell, 1997). Consequently, we use a paired *t*-test for testing the significance of the differences between the accuracies of the classification models as it is one of the most popular statistical tests in machine learning literature in spite of its shortcomings (Dietterich, 1996). Table IX lists

Table IX. Statistical significance of differences in performances of classification models

	p value					
Classification model (10-fold CV accuracy)	Model 1 (63.57%)	Model 2 (65.75%)	Model 3 (68.58%)	Model 4 (71.69%)	Model 5 (73.98%)	Model 6 (80.15%)
Model 2 (65.75%)	0.000044	—	—	—	—	—
Model 3 (68.58%)	—	0.000000	—	—	—	—
Model 4 (71.69%)	—	—	0.000000	—	—	—
Model 5 (73.98%)	—	—	—	0.000001	—	—
Model 6 (80.15%)	—	—	—	—	0.000000	—
Model 7 (81.84%)	—	—	—	—	—	0.000000

DOI: 10.1002/isaf

the results of the paired *t*-test on the 10 paired accuracies obtained from 10-fold CV for each of the six pairs of the SVM classification models (model 2 versus model 1; model 3 versus model 2; model 4 versus model 3; model 5 versus model 4; model 6 versus model 5; model 7 versus model 6).

As can be seen, the improvement in the performance of model 2 versus model 1 was found to be statistically significant at $p < 0.0001$ or lower. For model 2 versus model 1, the calculated *t*-value was 6.697221 with nine degrees of freedom and a *p*-value of 0.000044 (see Table IX) for an upper-tailed *t*-test. For model 3 versus model 2, the calculated *t*-value was 15.100643 with nine degrees of freedom and a *p*-value of 0.000000 for an upper-tailed *t*-test, and the improvement in the performance of model 3 versus model 2 was found to be statistically significant at $p < 0.0001$ or lower. For model 4 versus model 3, the calculated *t*-value was 31.649857 with nine degrees of freedom and a *p*-value of 0.000000 for an upper-tailed *t*-test, and the improvement in the performance of model 4 versus model 3 was found to be statistically significant at $p < 0.0001$ or lower. For model 5 versus model 4, the calculated *t*-value was 10.417465 with nine degrees of freedom and a *p*-value of 0.000001 for an upper-tailed *t*-test, and the improvement in the performance of model 5 versus model 4 was found to be statistically significant at $p < 0.0001$ or lower. For model 6 versus model 5, the calculated *t*-value was 79.580706 with nine degrees of freedom and a *p*-value of 0.000000 for an upper-tailed *t*-test, and the improvement in the performance of model 6 versus model 5 was found to be highly statistically significant at $p < 0.0001$ or lower. For model 7 versus model 6, the calculated *t*-value was 17.004234 with nine degrees of freedom and a *p*-value of 0.000000 for an upper-tailed *t*-test, and the improvement in the performance of model 7 versus model 6 was found to be statistically significant at $p < 0.0001$ or lower.

4.3.4. Top Discriminative Sentiment Features in Fraud Detection

We use the χ^2 statistical test to measure the importance of sentiment features in fraud detection. The greater the value of the χ^2 test statistic, the more useful the feature is in distinguishing fraudulent and truthful MD&As. Table X lists the 10 sentiment features ranked by their importance. Based on the χ^2 test scores, the top 10 most relevant sentiment features in distinguishing fraudulent and truthful MD&As arranged in descending order are adjectives (χ^2 value of 440.720), adverbs (χ^2 value of 321.59), negative words as per LM financial sentiment dictionary (χ^2 value of 277.64), adverbs modifying adjectives (χ^2 value of 258.19), positive words as per LM financial sentiment dictionary (χ^2 value of 194.85), superlative adjectives (χ^2 value of 170.22), verbs (χ^2 value of 164.07), affect emotion words as per LIWC dictionary (χ^2 value of 108.54), superlative adverbs (χ^2 value of 98.53) and subjective words as per MPQA subjectivity lexicon (χ^2 value of 51.01).

Table X. Top 10 sentiment features for distinguishing between fraudulent and truthful MD&As

Feature rank	Feature name	χ^2 statistic
1	Normalized frequency of adjectives	440.72
2	Normalized frequency of adverbs	321.59
3	Negative words as per LM financial sentiment dictionary	277.64
4	Normalized frequency of adverbs modifying adjectives	258.19
5	Positive words as per LM financial sentiment dictionary	194.85
6	Normalized frequency of superlative adjectives	170.22
7	Normalized frequency of verbs	164.07
8	Affect emotion words as per LIWC dictionary	108.54
9	Normalized frequency of superlative adverbs	98.53
10	Subjective words as per MPQA subjectivity lexicon	51.01

DOI: 10.1002/isaf

4.3.5. Further Investigation of Part-of-Speech Tags

The high informativeness of linguistically rich features such as POS tags in accurately recognizing fraudulent and truthful MD&As motivated us to do additional tests with POS features. We find that fraudulent MD&As, on average, have a higher percentage of adjectives and adverbs in contrast to truthful MD&As. In general, nouns tend to be more content dependent, whereas adjectives and adverbs tend to be strong indicators of subjectivity, and thus we do not include nouns in this analysis. We used information gain, which measures the decrease in entropy when the feature is present or absent, to discover adjectives and adverbs that have high information gain and are most useful for the classification. Table XI lists the top 15 adjectives and 15 adverbs based on information gain scores. We noted that several of these adverbs represent adverbs of doubt (e.g. approximately, partially, generally, likely) in contrast to adverbs that represent certainty (e.g. certainly, definitely). The high use of the adverb "also" suggests that the writer wants to highlight certain information and thus makes use of emphasis words to draw a user's attention to that information.

Next, we tested the statistical significance of the differences in the distribution of subjectivity features (namely, adjectives and adverbs) across the two corpora using the *t*-test. The results show that the sentiment style of fraudulent MD&As differs from truthful MD&As in certain types of noun modifiers (base form of adjectives) and verb modifiers (base form of adverbs). Table XII presents the results of the two-tailed *t*-test. As can be noted, the differences in the mean distribution of the POS-based subjectivity features (namely, adjectives and adverbs) in the fraudulent and truthful corpora were evaluated to be statistically significant, with $t = 10.3516$ with a value of $p < 0.0001$ for adjectives and with $t = 13.6894$ with a value of $p < 0.0001$ for adverbs.

Although sentiment polarity can be similar in some fraudulent and truthful MD&As, in general the two corpora are significantly different in their polarity. Compared with truthful MD&As, we find that negative and positive sentiment terms appear a lot more frequently in fraudulent MD&As. Using a two-tailed *t*-test, we noted that the average frequency of occurrence of positive sentiment words per MD&A document was three times more in the fraudulent MD&A corpora than in the truthful MD&A corpora, and the average frequency of occurrence of negative sentiment words per MD&A document was four times more in the fraudulent MD&A corpora than in the truthful MD&A corpora. Both of these differences between the two corpora were evaluated to be statistically significant by the *t*-test,

Table XI. Top 15 adjectives and adverbs for distinguishing fraudulent MD&As from truthful MD&As

Adjectives			Adverbs		
1. other	6. such	11. financial	1. also	6. adversely	11. respectively
2. fiscal	7. senior	12. global	2. approximately	7. accordingly	12. not
3. certain	8. fourth	13. observable	3. primarily	8. therefore	13. partially
4. prior	9. net	14. future	4. substantially	9. significantly	14. yet
5. due	10. additional	15. specific	5. additionally	10. only	15. generally

Table XII. Statistical significance of differences in the distribution of sentiment features between fraudulent and truthful MD&As

Feature name	*t*-statistic	Degrees of freedom	*p*-value (two-tailed)
Adjectives	10.3516	358	<0.0001
Adverbs	13.6894	358	<0.0001
Negative words as per LM financial sentiment dictionary	20.9151	358	<0.0001
Positive words as per LM financial sentiment dictionary	31.5286	358	<0.0001

DOI: 10.1002/isaf

with $t = 31.5286$ with a value of $p < 0.0001$ for positive sentiment words and with $t = 20.9151$ with a value of $p < 0.0001$ for negative sentiment words (see Table XII).

We further noted that use of intensity markers helped the model to effectively distinguish fraudulent MD&As from truthful MD&As that were previously misclassified by the models built without POS features. The reason for misclassification in the previous experiments was that these MD&A documents were similar with respect to polarity values and thus were confusing to the model, but the use of a particular intensity marker (i.e. "adverb modifying adjective") pattern resulted in reduced error rates, indicating that this pattern can better differentiate fraudulent MD&A text from truthful MD&A text when the two are similar in terms of sentiment polarity. Some examples of this "adverb adjective" pattern that occurred in MD&A corpus are "not successful", "primarily accountable", "approximately basic", "only reportable" and "ultimately subject". For example, Adelphia's 2001 MD&A was initially misclassified as truthful on the basis of polarity features only, but when we used this particular intensity feature in addition to other features the result was its correct classification as a fraudulent MD&A. A sample sentence extracted from Adelphia's MD&A (contained in the 2001 10-K filing) illustrates adverb adjective pattern:

> Equity in loss of joint ventures decreased in 2000 as compared to 1999, *primarily due* to the consolidation of Olympus (the adverb adjective pattern "**RB** primarily **JJ** due" is emphasized in the sample sentence). Another sample sentence from the same MD&A illustrates a further example of this pattern:
>
> The remaining increase was *primarily due* to increased net losses of Adelphia Business Solutions attributable to minority interests (the adverb adjective pattern "**RB** primarily **JJ** due" is emphasized in the sample sentence).

5. CONCLUSIONS

This study presents a novel approach to examine unstructured text contained in corporate disclosures. Using natural language processing techniques, we exploit sentiments that are expressed through text in the MD&A section of the annual reports. The results suggest that textual sentiment features (related to polarity, subjectivity and intensity) expressed in MD&A text can be used to identify companies that are at high risk of committing fraud as indicated by (1) more pronounced use of negative and positive sentiment, (2) higher proportion of subjective expressions than objective expressions, and (3) greater use of sentiment expressions that exhibit intensity of sentiment. The fraud sentiment classification model built with the most useful sentiment features accurately identified 81.84% of the MD&As and significantly outperformed all the other models built in this study. Six of the the top 10 most useful sentiment features were related to POS-based features and the remaining four were related to lexicon-based features, suggesting that a subset of features is sufficient to accurately distinguish between fraudulent and truthful MD&As.

The findings of this research lend support to the importance of examining the language of corporate disclosures for fraud detection in order to further advance this stream of fraud detection research in both academia and practice. These findings may be of particular interest to auditors, fraud examiners and regulators. For example, external auditors can use these findings to develop richer text analytic procedures to flag company filings that exhibit a higher proportion of sentiment markers identified in this study for deeper investigation. This is particularly important in light of the fact that auditors face high litigation risk associated with missed cases of fraud in addition to the high costs involved in modifying auditors' reports.

Use of sentiment features to detect language patterns in fraudulent and truthful documents is a difficult problem as corporate disclosures are typically considered to be dominated by factual information. The findings of this study can advance our understanding of sentiment in corporate disclosures that may be indicative of deception. Future research might investigate how use of intensifiers (diminishers) can strengthen

DOI: 10.1002/isaf

(weaken) the effect of the sentiment being expressed. Another direction for future research might examine other forms of corporate communication that may not be in a written format to gain additional insight into the use of sentiment to deceive users. One limitation of this study is that sentiment analysis is highly sensitive to the domain from which the training data are extracted. This is due to the fact that the same word can have a positive polarity in one domain and a negative polarity in another domain as sentiments are interpreted differently in different domains and contexts. As a result, the findings of this study may be more suitable to applications in the accounting domain and less suitable to applications in other domains.

ACKNOWLEDGEMENTS

We would like to thank Daniel E. O'Leary (editor), two anonymous reviewers and SET workshop participants of the 2013 American Accounting Association annual meeting for their constructive and useful comments on early drafts of this paper.

REFERENCES

Antweiler W, Frank MZ. 2004. Is all that talk just noise? The information content of Internet stock message boards. *Journal of Finance* **59**(3): 1259–1294.

Banfield A. 1982. Unspeakable Sentences. Routledge and Kegan Paul: Boston, MA.

Benamara F, Cesarano C, Picariello A, Reforgiato D, Subrahmanian VS. 2007. Sentiment analysis: adjectives and adverbs are better than adjectives alone. In Proceedings of the International Conference on Weblogs and Social Media. AAAI Press: Palo Alto, CA.

Bradley PS, Mangasarian OL, Street WN. 1998. Feature selection via mathematical programming. *INFORMS Journal on Computing* **10**(2): 209–217.

Bruce R, Wiebe J. 1999. Recognizing subjectivity: a case study of manual tagging. *Natural Language Engineering* **5**(2): 187–205.

Burgoon JK, Buller DB, Guerrero LK, Afifi WA, Feldman CM. 1996. Interpersonal deception: XII. Information management dimensions underlying deceptive and truthful messages. *Communication Monographs* **63**: 50–69.

Burgoon JK, Stoner GM, Bonito J, Dunbar NE. 2003. Trust and deception in mediated communication. In Proceedings of the 36th Annual Hawaii International Conference on System Sciences, 2003. IEEE Computer Press: Washington, DC.

Carlson JR, George JF, Burgoon JK, Adkins M, White CH. 2004. Deception in computer-mediated communication. *Group Decision and Negotiation* **13**: 5–28.

Cecchini M, Aytug H, Koehler GJ, Pathak P. 2010. Detecting management fraud in public companies. *Management Science* **56**(7): 1146–1160.

Chua C, Milosavljevic M, Curran JR. 2009. A sentiment detection engine for internet stock message boards. In *Proceedings of the Australasian Language Technology Association Workshop 2009*.

Churyk NT, Lee C, Clinton BD. 2009. Early detection of fraud: evidence from restatements. In Advances in Accounting Behavioral Research, Arnold V (ed), Vol. **12**. Emerald Group Publishing Ltd: Bingley, UK; 25–40.

Das SR, Chen MY. 2007. Yahoo! for Amazon: sentiment extraction from small talk on the Web. *Journal of Management Science* **53**(9): 1375–1388.

Dechow P, Ge W, Larson C, Sloan R. 2011. Predicting material accounting misstatements. *Contemporary Accounting Research* **28**(1): 17–82.

Devitt A, Ahmad K. 2007. Sentiment polarity identification in financial news: a cohesion based approach. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Association of Computational Linguistics: Stroudsburg, PA; 984–991.

Dietterich TG. 1996. Proper statistical tests for comparing supervised classification learning algorithms (Technical Report). Department of Computer Science, Oregon State University, Corvallis, OR.

Ding X, Lu B, Yu P. 2008. A holistic lexicon-based approach to opinion mining. In WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM: New York, NY; 231–240.

DOI: 10.1002/isaf

Dumais ST, Platt J, Heckerman D, Sahami M. 1998. Inductive learning algorithms and representations for text categorization. In CIKM '98 Proceedings of the Seventh International Conference on Information and Knowledge Management. ACM: New York, NY; 148–155.

Ekman P, Friesen WV. 1982. Felt, false, and miserable smiles. *Journal of Nonverbal Behavior* **6**(4): 238–252.

Esuli A, Sebastiani F. 2006. SENTIWORDNET: a publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*; 417–422.

Fludernik M. 1993. The Fictions of Language and the Languages of Fiction. Routledge: London.

Garcia D. 2013. Sentiment during recessions. *Journal of Finance* **68**(3): 1267–1300.

Glancy FH, Yadav SB. 2011. A computational model for financial reporting fraud detection. *Decision Support Systems* **50**(3): 595–601.

Goel S. 2014. Fraud detection and corporate filings. In Communication and Language Analysis in the Corporate World, Hart R (ed). Information Science Reference: Hershey, PA; 315–332.

Goel S, Gangolly J. 2012. Beyond the numbers: mining the annual reports for hidden cues indicative of financial statement fraud. *International Journal of Intelligent Systems in Accounting, Finance and Management* **19**: 75–89.

Goel S, Gangolly J, Faerman S, Uzuner O. 2010. Can linguistic predictors detect fraudulent financial filings? *Journal of Emerging Technologies in Accounting* **7**: 25–46.

Guo W. 2014. Executives' use of emotional language and investor reactions. In Communication and Language Analysis in the Corporate World, Hart R (ed). Information Science Reference: Hershey, PA; 198–215.

Hájek P, Olej V. 2013. Evaluating sentiment in annual reports for financial distress prediction using neural networks and support vector machines. In Engineering Applications of Neural Networks, Iliadis L, Papadopoulos H, Jayne C (eds), Vol. **384**Communications in Computer and Information Science. Springer: Berlin; 1–10.

Hájek P, Olej V, Renata M. 2013. Forecasting stock prices using sentiment information in annual reports – a neural network and support vector regression approach. *WSEAS Transactions on Business and Economics* **4**(10): 293–305.

Hancock JT. 2007. Digital deception: why, where and how people lie online. In The Oxford Handbook of Internet Psychology, Joinson AN, McKenna K, Postmes T, Reips U (eds). Oxford University Press: Oxford; 287–301.

Hancock JT, Woodworth M, Goorha S. 2010. See no evil: the effect of communication medium and motivation on deception detection. *Group Decision and Negotiation* **19**: 327–343.

Hart RP. 2000. Diction 5.0: The Text Analysis Program. Sage: Thousand Oaks, CA.

Hennes K, Leone A, Miller B. 2008. The importance of distinguishing errors from irregularities in restatement research: the case of restatements and CEO/CFO turnover. *The Accounting Review* **83**(6): 1487–1519.

Humpherys S, Moffitt KC, Burns MB, Burgoon JK, Felix WF. 2011. Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems* **50**: 585–594.

Joachims T. 1998. Text categorization with support vector machines: learning with many relevant features. In ECML'98 Proceedings of the 10th European Conference on Machine Learning, Nedellec C, Rouveirol C (eds). Springer-Verlag: London; 137–142.

Joachims T. 2001. A statistical learning model of text classification with support vector machines. In SIGIR '01 24th ACM/SIGIR International Conference on Research and Development in Information Retrieval, New Orleans, LA, USA — September 09–12, 2001. ACM: New York, NY; 128–136.

Li F. 2006. Do stock market investors understand the risk sentiment of corporate annual reports? Working paper, University of Michigan.

Liao XW, Cao DL, Tan SB, Liu Y, Ding GD, Cheng XQ. 2008. Combining language model with sentiment analysis for opinion retrieval of blog-post. In *The Fifteenth Text Retrieval Conference (TREC 2006) Proceedings*, Voorhees EM, Buckland LP (eds). NIST Special Publication: SP 500-272. NIST: Gaithersburg, MD. http://trec.nist.gov/pubs/trec15/papers/cas-ict.blog.final.pdf (accessed 22 April 2016).

Loughran T, McDonald B. 2011. When is a liability not a liability? Textual analysis, dictionaries and 10-Ks. *The Journal of Finance* **66**(1): 35–66.

Mitchell TM. 1997. Machine Learning. McGraw Hill: New York.

Newman ML, Pennebaker JW, Berry DS, Richards JM. 2003. Lying words: PREDICTING deception from linguistic styles. *Personality and Social Psychology Bulletin* **29**: 665–675.

O'Leary DE. 2011. Blog mining-review and extensions: "From each according to his opinion". *Decision Support Systems* **51**(4): 821–830.

Pennebaker JW, Booth RJ, Francis ME. 2007. Linguistic Inquiry and Word Count: LIWC 2007. LIWC: Austin, TX.

Platt JC. 1998. Fast training of support vector machines using sequential minimal optimization. In In *Advances in Kernel Methods: Support Vector Learning*, Scholkopf B, Burges C, Smola A (eds). MIT Press: Cambridge, MA.

DOI: 10.1002/isaf

Purda L, Skillicorn D. 2014. Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. *Contemporary Accounting Research* **32**: 1193–1223.

Roth D, Zelenko D. 1998. Part of speech tagging using a network of linear separators. In COLING–ACL, '98 Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Vol. **II**. Association for Computational Linguistics: Stroudsburg, PA; 1136–1142.

Skillicorn D, Purda L. 2012. Detecting fraud in financial reports. In Proceedings: 2012 European Intelligence and Security Informatics Conference, Memon N, Zeng D (eds). IEEE Computer Society: Odense, Denmark; 7–13.

Stone PJ, Dunphy DC, Smith MS, Ogilvie DM. 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press: Cambridge, MA.

Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics* **37**(2): 267–307.

Tetlock PC. 2007. Giving content to investor sentiment: the role of media in the stock market. *Journal of Finance* **62**: 1139–1168.

Thelwall M, Buckley K, Paltoglou G. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology* **63**(1): 163–173.

Tumarkin R, Whitelaw R. 2001. News or noise? Internet postings and stock prices. *Journal of Financial Analyst* **57**(3): 41–51.

Whissell CM. 1989. The dictionary of affect in language. In The Measurement of Emotions, Kellerman RP (ed). Academic Press: New York; 113–131.

Wiebe J. 1994. Tracking point of view in narrative. *Computational Linguistics* **20**(2): 233–287.

Wiebe J, Bruce R, O'Hara T. 1999. Development and use of a gold standard data set for subjectivity classifications. In ACL '99 Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics: Stroudsburg, PA; 246–253.

Wilson T, Wiebe J, Hoffman P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In HLT '05 Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics: Stroudsburg, PA; 347–354.

Witten IH, Frank E. 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. San Francisco, CA: Morgan Kaufmann.

Yu H, Yang J, Han J. 2003. Classifying large data sets using SVM with hierarchical clusters. In KDD '03 The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA — August 24–27, 2003. ACM: New York, NY; 306–315.

Zhou L, Burgoon JK, Nunamaker JF, Twitchell D. 2004. Automating linguistics-based cues for detecting deception in asynchronous computer-mediated communications. *Group Decision and Negotiation* **13**: 81–106.

DOI: 10.1002/isaf