Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2016 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

Summer 6-27-2016

FINANCIAL STATEMENT FRAUD DETECTION USING TEXT MINING: A SYSTEMIC FUNCTIONAL LINGUISTICS THEORY PERSPECTIVE

Wei Dong
City University of Hong Kong and USTC-CityU Joint Advanced Research Centre, dongwei3-c@my.cityu.edu.hk

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

Liang Liang
Hefei University of Technology, lliang@ustc.edu.cn

Follow this and additional works at: http://aisel.aisnet.org/pacis2016

Recommended Citation

Dong, Wei; Liao, Shaoyi; and Liang, Liang, "FINANCIAL STATEMENT FRAUD DETECTION USING TEXT MINING: A SYSTEMIC FUNCTIONAL LINGUISTICS THEORY PERSPECTIVE" (2016). *PACIS 2016 Proceedings*. 188.
http://aisel.aisnet.org/pacis2016/188

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2016 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

FINANCIAL STATEMENT FRAUD DETECTION USING TEXT MINING: A SYSTEMIC FUNCTIONAL LINGUISTICS THEORY PERSPECTIVE

Wei Dong, Department of Information Systems, City University of Hong Kong and USTC-CityU Joint Advanced Research Centre, Suzhou, China, dongwei3-c@my.cityu.edu.hk

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

Liang Liang, School of Management, Hefei University of Technology, Hefei, China, lliang@ustc.edu.cn

Abstract

Fraudulent financial information made by public companies not only cause significant financial loss to broad shareholders but also result in a great loss of confidence to capital market. Conventional auditing practices, which primarily focus on statistical analysis of structured financial ratios in auditing process, work not so well with the presence of misleading financial reports. This research tries to tap the power of huge amount of largely ignored textual contents in financial statements. With the theoretical guidance of Systemic Functional Linguistics theory (SFL), we develop a systematic text analytic framework for financial statement fraud detection. Seven information types, i.e., topics, opinions, emotions, modality, personal pronouns, writing style, and genres are identified based on ideational, interpersonal, and textual metafunctions in SFL. Under the analytic framework, Latent Dirichlet Allocation algorithm, computational linguistics, term frequency-inverse document frequency method, are integrated to create a synergy for extracting both word-level and document-level features. All these features serve as the input of Liblinear Support Vector Machine classifier. Finally, with application to detect fraud in 1610 firm-year samples from U.S. listed companies, the analytic framework makes a classification with average accuracy at 82.36% under ten-fold cross validation, much better than baseline method using financial ratios.

Keywords: Financial statement fraud, Fraud detection, Systemic Functional Linguistics theory, Text analytic framework.

1 INTRODUCTION

Financial statement fraud (FSF) is defined as “deliberate fraud committed by management that injures investors and creditors through misleading financial statements” (Elliott et al. 1980). Serious financial frauds such as Enron, WorldCom, and Tyco have not only bought substantial losses to shareholders but also shaken investors’ confidence in the integrity of stock market (Albrecht et al. 2008). More seriously, the number of fraudulent financial statements issued by listed companies in U.S. has increased incredibly in past ten years (Dechow et al. 2011). Effective and reliable FSF detection techniques are vital for preventing the devastating consequences of financial fraud.

Most financial fraud detection researches limit their investigations only to numerical data in financial statements (Humpherys et al. 2011). Due to deliberate concealment and/or accounting shenanigans, fraudulent financial data could hardly be distinguished from authentic data. Considering most of contents in financial statements, such as Form 10-K, are textual explanations for numerical data, researchers began to aware the value of this largely ignored textual information to detect financial fraud (Cecchini et al. 2010; Glancy et al. 2011; Humpherys et al. 2011). These researchers have verified the ability of the Management’s Discussion and Analysis (MD&A) section in financial statements for FSF detection. However, exist researches utilizing textual content in MD&A lack a systematic, holistic, and theoretical analytic framework to guide the fraud detection work and provide comprehensive textual features specific to FSF detection. It is actually the *raison d’etre* of this study.

In this research, Systemic Functional Linguistics theory (SFL) (Halliday et al. 2014) provides a useful theoretical foundation for the development of a text analytic framework to investigate the fraudulent behavior of top managements indicated in the language they use in MD&A section. SFL states that language is functional and making linguistic choices can help writer to achieve certain purpose. In turn, it will help us to understand the strategic language usage, especially the deceptive messages, of the writer. It has three metafunctions: ideational, interpersonal, and textual metafunction. In this research, the three metafunctions are conceptualized into seven information types: topics, opinions, emotions, modality, personal pronouns, writing style, and genres. Under the guidance of SFL theory, several commonly used theories, such as Interpersonal Deception Theory (IDT), Management Obfuscation Hypothesis (MOH), and three analysis methods, i.e., Latent Dirichlet Allocation (LDA), computational linguistics, term frequency-inverse document frequency (TF-IDF), are integrated and create a synergy for extracting word-level features and document-level features for all information types. All these features serve as the input of Liblinear Support Vector Machine (SVM) classifier. By examining 805 fraudulent firm-year samples and 805 non-fraudulent firm-year samples, the average testing accuracy can reach 82.36 percent, much better than the baseline method using financial ratios.

This research contributes to FSF detection from both theoretical and empirical angle. From a theoretical perspective, this research develops new feature selection process with solid theoretical guidelines of SFL theory. This study is the first to propose a systematic textual feature set for FSF detection. Second, this research introduces new constructs for FSF detection literature. Seven constructs, such as topics, opinions, emotions, modality, personal pronouns, writing style, and genres, are new in FSF detection area. Third, a text analytic framework that integrates LDA algorithm, computational linguistics, and TF-IDF methods is a new IT artifact for FSF detection (Hevner et al.

2004). From an empirical perspective, on one hand, the analytic framework provides a guideline for future text analysis in FSF detection. On the other hand, we demonstrate that textual features based FSF detection can be a complementary to existing financial ratios based techniques. This research will benefit financial governors and auditors in detecting fraud and protect the public's investments.

2 RESEARCH METHODOLOGY

Management's Discussion and Analysis refers to a section in financial statements or Form 10-K, which is written by top managements for providing investors with a sense of how the business performed in the past, its current financial condition as well as projections of future outlook. Once a company faces market-driven pressures due to predicament, the managements have the incentive to commit fraud since poor operation performance and falling stock value will result in low compensation comes from stock options (Hake 2005). They will create various misleading statements, including, but not limited to, presenting misleading optimistic past performance, providing ambiguous information about the current health, too positively stating future outlook. Therefore, this research develops a novel text analytic framework, guided by SFL theory, to investigate the fraudulent behavior of top managements indicated in the language they use in MD&A section. Halliday et al. (2014) analyzed language into three interrelated metafunctions: ideational, interpersonal, and textual metafunction. Each metafunction will be concrete into different information types thereinafter. Then features for each information type are extracted from textual contents of MD&A section. The processes of feature extraction and classification constitute the analytic framework (shown in Figure 1). The framework can also serve as a system development framework for developing FSF detection systems in practice.

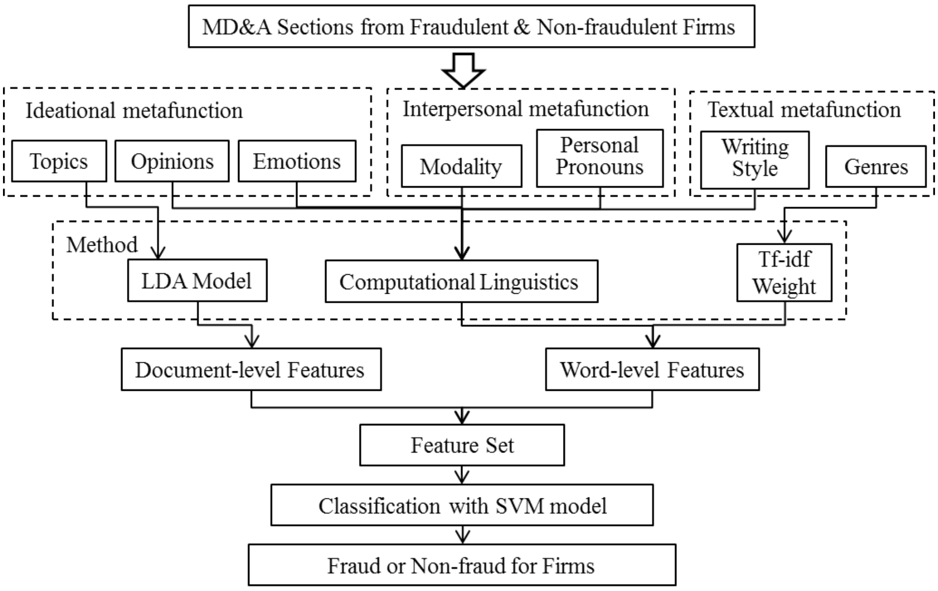

Figure 1. Text Analytic and System Development Framework for FSF Detection.

2.1 Identification of Information Types from Metafunctions of SFL theory

The ideational metafunction tends to make meanings about the world around and inside us. It can be represented by topics, opinions, and emotions (Abbasi et al. 2008). Topic, in linguistics, also known as theme, of a sentence is what is being talked about. According to Reality Monitoring and Criteria-based content analysis (Zhou et al. 2004), a statement derived from truth differs in content and quality from that been made up. By control the topics discussed, deceivers can make a strategic use of deceptive language in text. Opinions are sentiment polarities (e.g., positive, neutral, and negative) about a particular entity (Pang et al. 2008). Managers engaged in fraud are more likely to include more positive words and portray their company operation and forward looking in much more positive light. Emotions consist of various affects such as happiness, sadness, horror, and anger (Abbasi et al. 2008). If the fraudulent behaviour is reflected in the language use, the text could be filled with more words reflecting negative emotion (e.g., hate, sad, anger) (Newman et al. 2003).

The interpersonal metafunction of a written language represents the way the writer and the readers interact, and the use of language to establish and maintain relations with them. Writers use language to express attitude towards the subject matters and to influence readers' behaviour. Information types such as modality and personal pronoun identified by Halliday et al. (2014) are adopted for representing the interpersonal metafunction. Modality shows writer's judgment of the validity of the proposition. For example, modal word "must" represents a strong modal commitment, which signals a high degree of certainty about the validity of a proposition; On the contrary, "could" represents a low value judgment. Hence, we expect that textual languages in MD&A are more likely to contain more weak-modality words and less strong-modality words to undertake uncertainty manipulation when managers are found with fraudulent behaviours. In addition, deceivers who want to disassociate themselves with responsibility for misstatements in financial statements will decrease self-reference and active voice usage as indicated by Interpersonal Deception Theory (IDT) (Buller et al. 1996). We, therefore, also expect MD&A from fraudulent companies to include more non-immediate language by using less self-reference, and more other reference, group reference, and passive voice than truth-teller.

The textual metafunction refers to how language is organized and structured to create a coherent and continuous flow of information. Two information types, i.e., writing style, genres, conceptualized in Argamon et al. (2007) are considered in this research. Writing style is based on the literary choices a writer makes, which can be a reflection of context (who, what, when, why, where) (Abbasi et al. 2008). Deceptive statements must be presented in a style that pretends to be authentic and sincere. Goel et al. (2012) stated that writing style changes when a company is committing fraud. Genres represent how writers typically use language to respond to recurring situations (Hyland 2004). As an identifiable genre for business communication (Merkl-Davies et al. 2007), narratives in financial statement are able to contain various sub-genres. Differences in genres are expected to be identified between fraudulent and legitimate company narratives.

2.2 Measurement of information types

In this section, text mining techniques are used to extract features for all information types in three metafunctions of SFL theory.

In ideational metafunction, semantic topics of financial statements are extracted using widely used topic model Latent Dirichlet Allocation (LDA) (Blei et al. 2003). It assumes that each individual document is a mixture of various topics and each topic is a probability distribution over a group of words. By running LDA model, latent topics in the document can be extracted automatically. Each document is explicitly represented by a low-dimension vector of topic probabilities at last. For opinion information type, it is measured by ratios of both positive and negative sentiment words using sentiment words dictionary in financial domain created by Loughran et al. (2011). In this study, word categories “positive emotion”, “negative emotion”, “anxiety”, “anger”, and “sadness” from Linguistic Inquiry and Word Count (LIWC, Pennebaker et al. (2001)) dictionary are adopted to compute the corresponding word ratios for representing emotions in MD&A. Features and explanations for ideational metafunction are summarized in Table 1.

In interpersonal metafunction, modality is first measured by ratio of modal verbs. Besides, common modal words, such as “always”, “never”, “possibly”, are also counted. Hence, we adopt the strong and weak modal words classification in Loughran et al. (2011) as another two features. The personal pronoun information type is measured by ratios of self-references, group-references and other references generated from LIWC dictionary. Since passive voice indirectly signals personal pronoun, ratio of passive verbs is also considered. Table 1 explains details about features of this metafunction.

Considering writing style in textual metafunction, it is first measured in three aspects, i.e., complexity, pausality, and expressivity (Zhou et al. 2004). Complexity is measured by average number of clauses, average sentence length, average word length, average length of noun phrase (Zhou et al. 2004). Pausality is measured by the ratio of punctuation marks (Humpherys et al. 2011). In terms of expressivity, it is indicated by emotiveness (Humpherys et al. 2011). In addition, according to Management Obfuscation Hypothesis (MOH) (Bloomfield 2002), management from bad performance or fraud perpetrating company tends to obfuscate information in MD&A section. One direct way to increase obfuscation is to reduce the readability of the text. As suggested by Li (2008), we add Fog index and the logarithm length of document to measure readability of MD&A. All features for writing style are explained in Table 1. Enlighten by genre analysis framework in Rutherford (2005), text genre can be analyzed using word frequency. Hence, term frequency-inverse document frequency (TF-IDF) method is adopted to count word frequency in whole corpus. Noted that TF-IDF word weight vector is always of high dimension, only significant features at 5% level by a paired sample T-test are selected.

So far we have conceptualized SFL theory into seven information types and identified related features for each information type. The document-topic vector for topic information type is document-level features while features for other information types are word-level features.

Metafunction	Information type	Features	Feature explanation
Ideational	Topics	Document-topic vector	Low dimension document topic distribution extracted by LDA model.
	Opinions	Ratio of positive sentiment words	Total number of positive sentiment words divided by total number of words[a].
		Ratio of negative sentiment words	Total number of negative sentiment words divided by total number of words.
	Emotions	Ratio of positive emotion words	Total number of positive emotion words divided by total number of words.
		Ratio of negative emotion words	Total number of negative emotion words divided by total number of words.

		Ratio of anxiety words	Total number of anxiety words divided by total number of words.
		Ratio of anger words	Total number of anger words divided by total number of words.
		Ratio of sadness words	Total number of sadness words divided by total number of words.
Interpersonal	Modality	Ratio of modal verbs	Number of modal verbs divided by the total number of verbs.
		Ratio of strong modal words	Number of strong modal words divided by total number of verbs.
		Ratio of weak modal words	Number of weak modal words divided by total number of verbs.
	Personal Pronoun	Ratio of self-references	Total number of first person singular pronouns divided by total number of verbs.
		Ratio of group references	Total number of first person plural pronouns divided by total number of verbs.
		Ratio of other references	Total number of all other person singular or plural pronouns divided by total number of verbs.
		Ratio of passive verbs	Number of passive verbs divided by total number of verbs.
Textual	Writing Style	Average number of clauses	Total number of clauses divided by total number of sentences.
		Average sentence length	Total number of words divided by total number of sentences.
		Average word length	Total number of characters divided by total number of words.
		Average length of noun phrase	Total number of words in noun phrases divided by total number of noun phrases.
		Ratio of punctuation marks	Number of punctuation marks divided by total number of sentences.
		Emotiveness	Total number of adjectives and adverbs divided by total number of nouns and verbs.
		Fog index	(Average sentence length + percent of complex words[b]) ×0.4,
		Logarithm length of document	Log (the number of words in documents)
	Genre	Word frequency	Significant TF-IDF weight of words under paired sample T-test.

a. Total number of words are amount of words ignoring articles (a, an, the), the same hereinafter.

b. Percent of complex words = the number of words with three syllables or more divided by total number of words.

Table 1. Features Extracted for Financial Statement Fraud Detection.

2.3 Text classification

Following a supervised learning paradigm, SVM model is adopted for classifying financial statements of fraudulent and non-fraudulent firms due to its competitive advantages compared with other classifiers (Cecchini et al. 2010). Especially, a special SVM model, i.e., Liblinear (Fan et al. 2008), which is very efficient on large-scale feature set, is adopted. The dependent variable is a binary variable, indicating whether a financial statement for a fiscal year is related to financial fraud or not. Standard evaluation metrics such as accuracy, precision, recall, F1 score, false positive rate (FPR), and false negative rate (FNR) are used to evaluate the performance of this analytic framework.

3 DATA COLLECTION

In this study, we select the fraudulent financial statement cases from companies in American capital market to test the feasibility and performance of the proposed analytic framework. U.S. Securities and Exchange Commission (SEC) has been issuing AAERs, since 1982, to investigate a company, or other related parties for alleged accounting misconduct. We utilize these AAERs to screen companies issuing fraudulent financial statements. Ultimately, we find 319 distinct fraudulent firms with 805 fraud-year samples during the period from 17 May 1982 to 31 December 2014. For each company in fraudulent samples, we match it with a control sample, a non-fraudulent company, for classification purpose. Non-fraudulent sample are matched with the fraudulent sample directly by using COMPUSTAT on the basis of year, size, and industry. Therefore, there are 805 fraud-year samples and 805 nonfraud-year samples in this research.

4 EMPIRICAL TESTING

The MD&A section of each financial statement for all firm-year samples is identified and extracted into individual text file. Stop words are removed according to the stop words list created by Loughran et al. (2011), which is mainly for financial materials.

4.1 Test of the Proposed Analytic Framework for Classification

In this research, LDA model is used in the discriminative framework, in which document-topic vectors for all firm-year samples are estimated all at once without referencing to their true class labels. The number of topics is set as one hundred for simplicity. Number count and ratio computation for features in opinions, emotions, modality, personal pronouns, and writing style are identified for all samples prior to classification as well. Then by adopting a ten-fold validation approach, nine-tenths of 1610 firm-year samples are used to train (or build) the SVM prediction model and the other one-tenth samples are remained for testing the performance of the model built in each fold. Note that TF-IDF weights vectors for genres are only computed using words in training samples not considering testing samples. In other words, TF-IDF weights are computed ten times in ten-fold cross validation.

As shown in Table 2, the analytic framework can classify training samples with average accuracy, precision, recall, F1 score more than 99 percent, and average FPR and FNR almost zero. Testing samples are predicted with average accuracy at 82.36 percent and average precision, recall, and F1 score are all more than 81 percent. It indicates that the performance of proposed feature set for FSF detection ranks among the top in literature.

4.2 Comparison with Baseline Method

For comparison purpose, accounting method using financial ratios to detect financial fraud discussed in Abbasi et al. (2012) is selected as baseline method. Based on 12 seed financial ratios, Abbasi et al. (2012) created overall 84 financial ratios finally. These financial ratios for each firm-year sample are computed based on data retrieved from COMPUSTAT database. Following a same SVM model under ten-fold validation, the average training accuracy is 65.97 percent and F1 score is 62.66 percent. The average testing accuracy is 52.29 percent and F1 score is 59.89 percent. Detail results are shown in Table 2. The comparably weak performance of baseline method, to some extent, is attributed to missing values in calculating financial ratios. Nevertheless, it still clearly shows that classification using proposed analytic framework is much better than the baseline method.

4.3 Test of Classification Performance with Combined Feature Set

Furthermore, we combine the proposed feature set and these 84 financial ratios together to examine the classification performance for these firm-year samples. The training performance of combined feature set, i.e., no classification errors, is better than that only using analytic framework. The average testing accuracy, recall, F1 score, and FPR of combination method are better than that only using analytic framework while average precision and FNR are a little bit worse. A substantial improvement of the baseline method demonstrates the features developed by analytic framework can be complementary to conventional financial ratios for FSF detection.

		Average accuracy	Average precision	Average recall	Average F1 score	Average FPR	Average FNR
Analytic framework	Training	99.94	99.92	99.96	99.94	0.08	0.04
	Testing	82.36	81.48	86.23	83.00	20.53	13.77
Baseline method	Training	65.97	67.09	58.94	62.66	28.83	41.06
	Testing	52.29	66.00	57.84	59.89	32.87	42.16
Combination feature set	Training	100	100	100	100	0	0
	Testing	82.49	78.33	92.06	84.37	7.94	27.27

Table 2. Comparison of classification results.

5 CONCLUSIONS

In this research, we develop a text analytic framework, including features extraction and text classification, for FSF detection. The major contribution of this research is the feature set developed under the guidance of SFL theory. Using the proposed feature set, we obtain average prediction accuracy at 82.36 percent, much better than baseline method. We have also verified that the proposed feature set can be complementary to existing accounting method using financial ratios. With the help of financial fraud detection method using combined feature set, investors will make informed investment decisions, auditors will better assess the fraud risk of a focal firm, and regulators will allocate limited resources to investigate only most suspicious firms.

There are also some works left for future. First, except for a simple comparison between the analytic framework and benchmark method using financial ratios, other text analysis based detection methods considering textual contents in MD&A section are to be compared, such as methods in Humpherys et al. (2011) and Glancy et al. (2011). Second, classification results are needed to be compared among different machine learning classifiers. Third, more non-fraudulent firm-year samples are to be considered in classification. In this study, we choose a balanced sample dataset. However, given that fraud occurs less than 1% of the time (Beneish 1999), a balanced data set is far from mirroring reality. Favorable results found using proposed text analytic framework are to be tested on non-balanced samples in future research. At last, effect of the number of topics on the classification performance is to be studied instead of manually set one in current research.

ACKNOWLEDGMENTS

This research was supported by a Hong Kong GRF grant (no 193213), and a development grant (no ZDSYS20140509155229805) together with a fundamental research grant (no JCYJ20140630144136828) from Shenzhen Science and Innovation Commission.

References

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

Abbasi, A., and Chen, H. (2008). CyberGate: A design framework and system for text analysis of computer-mediated communication. MIS Quarterly, 32 (4), 811-837.

Albrecht, W. S., Albrecht, C., and Albrecht, C. C. (2008). Current trends in fraud and its detection. Information Security Journal: A Global Perspective, 17 (1), 2-12.

Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., and Levitan, S. (2007). Stylistic text classification using functional lexical features. Journal of the American Society for Information Science and Technology, 58 (6), 802-822.

Beneish, M. D. (1999). The detection of earnings manipulation. Financial Analysts Journal, 55 (5), 24-36.

Blei, D. M., Ng, A. Y., Jordan, M. I., and Lafferty, J. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3 (4/5), 993-1022.

Bloomfield, R. J. (2002). The "Incomplete Revelation Hypothesis" and financial reporting. Accounting Horizons, 16 (3), 233-243.

Buller, D. B., and Burgoon, J. K. (1996). Interpersonal deception theory. Communication theory, 6 (3), 203-242.

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

Dechow, P. M., Ge, W., Larson, C. R., and Sloan, R. G. (2011). Predicting material accounting misstatements. Contemporary Accounting Research, 28 (1), 17-82.

Elliott, R. K., and Willingham, J. J. (1980). Management fraud: Detection and deterrence. (Petrocelli Books New York.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research, 9, 1871-1874.

Glancy, F. H., and Yadav, S. B. (2011). A computational model for financial reporting fraud detection. Decision Support Systems, 50 (3), 595-601.

Goel, S., and Gangolly, J. (2012). Beyond the numbers: Mining the annual reports for hidden cues indicative of financial statement fraud. Intelligent Systems in Accounting, Finance and Management, 19 (2), 75-89.

Hake, E. R. (2005). Financial illusion: Accounting for profits in an enron world. Journal of Economic Issues (Association for Evolutionary Economics), 39 (3), 595-611.

Halliday, M., Matthiessen, C. M., and Matthiessen, C. (2014). An introduction to functional grammar. Routledge.

Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28 (1), 75-105.

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

Hyland, K. (2004). Genre and second language writing. University of Michigan Press.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. Journal of Accounting and Economics, 45 (2–3), 221-247.

Loughran, T. I. M., and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66 (1), 35-65.

Merkl-Davies, D. M., and Brennan, N. (2007). Discretionary disclosure strategies in corporate narratives: incremental information or impression management. Journal of Accounting Literature, 26, 116-196.

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

Pang, B., and Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2 (1-2), 1-135.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic Inquiry and Word Count (LIWC): A computerized text analysis program. Mahwah (NJ), 7.

Rutherford, B. A. (2005). Genre analysis of corporate annual report narratives a corpus linguistics–based approach. Journal of Business Communication, 42 (4), 349-378.

TEO, P. (2000). Racism in the news: a critical discourse analysis of news reporting in two australian newspapers. Discourse & Society, 11 (1), 7-49.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., and Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. Group decision and negotiation, 13 (1), 81-106.