Financial Statement Fraud Detection Using Text Mining: A Systemic Functional Linguistics Theory Perspective

Wei Dong[a,b], Stephen Shaoyi Liao[a], Liang Liang[c]
a. Department of Information Systems, City University of Hong Kong
b. School of Management, University of Science and Technology of China
c. School of Management, Hefei University of Technology

1. Introduction

Financial statement fraud is defined as "deliberate fraud committed by management that injures investors and creditors through misleading financial statements" (Elliott et al. 1980). It is a serious problem worldwide. Most financial fraud detection researches limit their investigations only to numerical data in financial statements. However, 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 detecting financial statement fraud. However, research utilizing textual content in MD&A lacks a systematic, holistic, and theoretical analytic framework to guide the fraud detection work and provide comprehensive textual features specific to statement fraud 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 and analytic framework to develop feature set for financial statement fraud detection. SFL explores how language is used in social contexts to achieve particular goals. On one hand, SFL places higher importance on language function (what it is used for) than on language structure (how it is composed). It is well suited for the textual analysis of MD&A section since all MD&As have unified structure and structure cannot be used to distinguish deceptive and true statements in MD&A. On the other hand, fraud in MD&A section is a purposeful, strategic deception perpetrated by managements using deceptive and misleading languages. It is believed that SFL is a potent tool for uncovering the stratagems that fraud perpetrator use to convince others of their view-points. By studying the language difference, it provides an opportunity to distinguish fraudulent and non-fraudulent text.

Under the guidance of SFL theory, 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 both word-level features and document-level features. All these features serve as the input of 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 under ten-fold cross validation, much better than the baseline method using numerical data.

2. Research Methodology

2.1 Metafunctions of Systemic Functional Linguistics theory

In order to understand the meaning of text, Halliday et al. (2014) analyzed language into three interrelated metafunctions: ideational, interpersonal, and textual metafunction. Ideational metafunction tend to make meanings about the world around and inside us. Interpersonal metafunction refers to language as a medium for interaction and as means for creating and maintaining our interpersonal relations. These two metafunctions in discourse are organized and weaved via the textual metafunction, which refers to how information is organized and presented and to create a coherent flow of discourse. Different information types are identified for each metafunction as follows.

The ideational metafunction can be represented by topics, opinions, and emotions (Abbasi et al. 2008). Topic, in linguistics, also referred to as theme, of a sentence is what is being talked about. According to Reality Monitoring (RM) and Criteria-based content analysis (CBCA) (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 than that of non-fraudulent companies. Emotions consist of various affects such as happiness, sadness, horror, and anger (Abbasi et al. 2008). If the fraud behavior 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' behavior. 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 words "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 fraud behaviors. In addition, if deceivers who want to disassociate themselves with responsibility for misstatements in financial statements will decrease self-reference and active voice usage. 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, structured to create a coherent and continuous flow of information (Halliday et al. 2014). 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 shown 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. Features and explanations for interpersonal metafunction are shown in Table 1.

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), 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 annual report readability. All features for writing style are explained in Table 1. In the light of genre analysis framework proposed by Rutherford (2005), text genre can be analyzed using word frequency based on corpus linguistics. In this study, we adopt term frequency-inverse document frequency (TF-IDF) method to count word frequency in whole corpus. Noted that TF-IDF word weight vector is always of high dimension, only features under 0.05 significant level by a paired sample T-test are selected.

Table 1. Features Extracted for Financial Statement Fraud Detection

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.

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.

2.3 Text classification

Following a supervised learning paradigm, SVM model is adopted for classifying financial statements of fraudulent and non-fraudulent firms. Especially, a special SVM model, i.e., Liblinear (Fan et al. 2008), which is very efficient on large-scale data 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.

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 feature set. 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 involving 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 sample, we match it with a control sample, a non-fraudulent company, for classification purpose. Therefore, there are 805 fraud-year samples and 805 nonfraud-year samples in this research.

4. Empirical results

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 proposed feature set 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 proposed feature set 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 financial statement fraud 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 are computed based on data retrieved from COMPUSTAT database. Limited by missing values in calculating financial ratios, only 297 fraudulent firm-years and 297 legitimate firm-years are remained in baseline method. 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. It's clear that classification using proposed feature set 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 297 fraudulent firm-year and 297 legitimate firm-year samples. The training performance of combined feature set, i.e., no classification errors, is better than that only using proposed feature set. The average testing accuracy, recall, F1 score, and FPR of combination method are better than that only using proposed feature set while average precision and FNR are a little bit worse. A substantial improvement of the baseline method

demonstrates the proposed feature set can be complementary to conventional financial ratios for financial statement fraud detection.

Table 2. Comparison of classification results

		Average accuracy	Average precision	Average recall	Average F1 score	Average FPR	Average FNR
Proposed feature set	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

5. Conclusions

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.

References:

Abbasi, et al., 2012. Metafraud: a meta-learning framework for detecting financial fraud, *Mis Quarterly* 36(4), 1293-1327.

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

Argamon, et al., 2007. Stylistic text classification using functional lexical features, *Journal of the American Society for Information Science and Technology* 58(6), 802-822.

Blei, et al., 2003. Latent Dirichlet Allocation, *Journal of Machine Learning Research* 3(4), 993-1022.

Cecchini, et al., 2010. Making words work: Using financial text as a predictor of financial events, *Decision Support Systems* 50(1), 164-175.

Elliott, et al., 1980. Management fraud: Detection and deterrence, *Petrocelli Books New York*.

Fan, et al., 2008. LIBLINEAR: A library for large linear classification, *Journal of Machine Learning Research* (9), 1871-1874.

Glancy, et al., 2011. A computational model for financial reporting fraud detection, *Decision Support Systems* 50(3), 595-601.

Goel, et al., 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.

Halliday, et al., 2014. An introduction to functional grammar, *Routledge*.

Humpherys, et al., 2011. Identification of fraudulent financial statements using linguistic credibility analysis, *Decision Support Systems* 50(3), 585-594.

Hyland, 2004. Genre and second language writing, *University of Michigan Press*.

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

Loughran, et al., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, *The Journal of Finance* 66(1), 35-65.

Merkl-Davies, et al., 2007. Discretionary disclosure strategies in corporate narratives: Incremental information or impression management?, *Journal of Accounting Literature* (26), 116-196.

Newman, et al., 2003. Lying words: Predicting deception from linguistic styles, *Personality and Social Psychology Bulletin* 29(5), 665-675.

Pang, et al., 2008. Opinion mining and sentiment analysis, *Foundations and trends in information retrieval* 2(1-2), 1-135.

Pennebaker, et al., 2001. Linguistic Inquiry and Word Count (LIWC): A computerized text analysis program, *Mahwah (NJ)* (7).

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

Zhou, et al., 2004. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications, *Group decision and negotiation* 13(1), 81-106.