Exploring the Information Contents of Risk Factors in SEC Form 10-K: A Multi-Label Text Classification Application

Article *in* SSRN Electronic Journal · October 2010

DOI: 10.2139/ssrn.1784527

CITATIONS
7

READS
1,105

1 author:

Ke-Wei Huang
National University of Singapore
48 PUBLICATIONS **410** CITATIONS

SEE PROFILE

EXPLORING THE INFORMATION CONTENTS OF RISK FACTORS IN SEC FORM 10-K: A MULTI-LABEL TEXT CLASSIFICATION APPLICATION

Ke-Wei Huang[1]

National University of Singapore

huangkw@comp.nus.edu.sg

Abstract

This study develops, implements, and evaluates a multi-label text classification algorithm that extracts textual information from the annual reports of all publicly listed USA companies. Specifically, the proposed system can automatically identify 25 types of frequently mentioned risk factors in a section called "Item 1A. Risk Factors" in SEC Form 10-K. The true positive rate on the training set with 3,153 risk factors is 80.65 percent while the false positive rate is 12.67 percent. This system is applied to extract risk factors in 10-Ks of most USA companies from 2006 to 2010. By the first-differencing panel data regression, this study shows that the extracted risk factors, the associated risk factor orderings, and the number of risk factors indeed provide additional explanation power for the target firm's risk measures, annual stock returns, and key financial ratios.

Keywords: *Text classification; text mining; multi-label classification; risk factors; annual reports; financial statement analysis.*

[1] A similar paper by the same author is "*A Multi-Label Text Classification Algorithm for Labeling Risk Factors in SEC Form 10-K*" (ACM Transactions on Management Information Systems, forthcoming). This article focuses on the performance comparions of the proposed algorithm and other existing algorithms on classifying risk factors in SEC Form 10-K.

EXPLORING THE INFORMATION CONTENTS OF RISK FACTORS IN SEC FORM 10-K: A MULTI-LABEL TEXT CLASSIFICATION APPLICATION

1. INTRODUCTION

It is widely recognized that corporate annual reports play a key role in financial markets. Corporate annual reports are regarded as one of the most important sources of information about a company. Disclosures in annual reports enable banks, institutional investors of stocks or corporate bonds, central banks, government's economic departments, upstream suppliers or downstream distributors of the target firm to make more accurate, critical decisions based on more precise evaluation of the risk, return, and operational performance of the target companies. Due to the increasingly huge number of publicly listed companies and the abundant textual information provided in the annual reports, computerized text analysis should be able to play a more valuable role to alleviate the information overloading problem in this context.

This study focuses on extracting risk factors from SEC Form 10-K. The Securities and Exchange Commission (SEC) requires all publicly-held U.S. companies to file reports disclosing their financial condition, results of operations and any other information that is of significance to investors. Form 10-K is the report filed annually by U.S. companies 90 days after the end of each fiscal year. 10-Ks are much more detailed than the annual reports sent to shareholders. Most pages of 10-K include unstructured, qualitative descriptions about various aspects of the target company. However, existing academic studies or decision support systems utilized mostly the numbers in annual reports. The reason is obvious: textual information is hard to be quantified and requires domain experts, such as stock analysts or researchers, subjectively digest and interpret

that information. It is natural to expect that the value of those unstructured information is tremendous both for research opportunities and for building investment or financial analytics applications.

Recent developments in text mining have enabled researchers to extract much more information from textual data available on the Internet. Rigorous studies by computational text analysis started to emerge in accounting and finance, but are still at the nascent stage. Existing studies in accounting or finance typically use basic features provided by content analysis software packages while few studies have applied text classification (See Table 1). This provides an excellent opportunity for information systems researchers to contribute to other disciplines.

This study demonstrates a novel way to extract unexplored yet valuable information from SEC 10-K. Specifically, a new multi-label text classification algorithm is proposed and is implemented to extract the meaning of "risk factors" reported in Section 1A of SEC 10-K forms. The definition and examples of "risk factors" are provided in Section 3. In plain language, "risk factors" are bullets of short sentences that describe what could go wrong, what are likely external negative factors, what are possible future failures to meet any obligations, and any other risks that should be disclosed to adequately warn investors. The proposed algorithm can automatically label 25 types of frequently mentioned risk factors with 80.65% true positive rate and 12.67% false positive rate on the training set.

This study also evaluates the information content of those 25 risk factors by regressing risk factors on ten measures of risks, stock returns, and financial ratios of most USA publicly listed companies from 2005 to 2010. The main econometric model is the first-differencing model, which is equivalent to the fixed effect panel regression. Results suggest that these 25 risk factors, the ordering of risk factors, and the number of risk

factors reported in one 10-K indeed contain significant explanatory power for ten frequently-used financial performance measures.

The remainder of this paper is organized as follows. Section 2 discusses related research. Section 3 details the data collection issues for the text classification task, including the definitions and examples of risk factors. Section 4 describes a new algorithm and its performance. Section 5 reports the empirical results of using a panel regression to examine the informativeness of the extracted risk factors. Section 6 concludes.

2. LITERATURE

The present study is most related to quantifying textual information in 10-K or other SEC forms. Without computer's help, researchers can only manually code a small number of samples of SEC forms for research. This approach has two shortcomings: the sample size is limited and the consistency of the coding is subjective and questionable. As a result, "content analysis" without computer's help has not been widely adopted in economics, finance, or accounting literature.

2.1 Computational content analysis in accounting and finance

Recently, accounting researchers started to apply various computational methods, especially content analysis software packages, to summarize textual information in financial statements for accounting studies. In an award winning paper in accounting, Li (2008) studied the relationship between corporate earnings and the readability of 10-K filings. The author used two variables to measure readability: (1) the Fog Index which is computed based on complex words with three or more syllabus and average sentence

length, and (2) the length of the 10-K itself. The author found that companies with less readable (higher Fog Index and longer in length) 10-Ks have lower earnings.

After this pioneering paper, a growing number of accounting or finance studies started to use "tone", "sentiment", or "readability" computed from content analysis software packages to investigate various issues. For example, Feldman et al. (2009) classify words into positive and negative categories to measure the tone change in the management discussion and analysis (MD&A) section in 10-Q and 10-K. The authors find that stock market reactions around the SEC filing are significantly associated with the tone change of the MD&A section, even after controlling for accruals and earnings surprises. Kothari et al. (2009) find that when content analysis indicates favorable disclosures in all media channels, the firm's risk declines significantly. Loughran and McDonald (2010) improve the readability used in Li (2008) for 10-K.

There is a similar trend of using text analysis in the finance literature, Tetlock (2007) studies the relationship between the stock market and the content from a Wall Street Journal (WSJ) column named "Abreast of the Market". The independent variable is a sentiment index calculated by the number of positive words and negative words categorized by General Inquirer (GI), a well-known content analysis program originally used for psychologists. This study finds that media pessimism induces downward pressure on market prices. Following Tetlock (2007), Tetlock et al. (2008) quantified financial news stories by the same measure to predict firms' accounting earnings and stock returns. Their findings suggest that negative words are more influential than positive words and can be utilized to forecast low firm earnings. A recent paper by Loughran and McDonald (2009) shows that negative words categorized by GI may not have actual negative meaning in 10-K. A new word classification is developed and this study shows that negative words by the new classification are more informative than those in GI.

2.2 Text classification studies in accounting and finance

Applying text classification (rather than using content analysis software) to study finance or accounting issues is still at its nascent stage. Earlier studies focused on stock price prediction. Antweiler and Frank (2004) classify messages posted on Yahoo! Finance by using Naïve Bayes and Support Vector Machine. Individual message was classified as bullish, bearish, or neutral. This study finds that messages can help to predict market volatility but not expected return. Similarly, Das and Chen (2007) classify Yahoo! Finance messages about technology companies into three groups: bullish, bearish and neutral by the voting of 5 classification algorithms. The overall evidence suggests that market activity is related to small investor sentiment and message board activity.

Recent studies focused more on extracting information from SEC forms. Li (2010) utilizes Naïve Bayes classification to classify the tone of "forward-looking statements" and found that a change in the tone imply a future change in the performance of the company. Balakrishnan et al. (2010) classify each 10-K into three classes: out-performing, average and under-performing by the historical performance. Analyses show that this model captures information not contained in document-level features of clarity, tone and risk sentiment but does not provide information incremental to firm size, market-to-book and momentum. Mangen and Durnev (2010) study the tone in restatement announcements by content analysis software. The authors find that restatement tone affects restating firms' and their competitors' abnormal returns. Hanley and Hoberg (2010) study the information content of IPO prospectuses and IPO pricing. The authors find that MD&A most contributes to the informativeness of the prospectus.

Table 1 Summary of the Text Analysis Methods in Accounting and Finance Literature

Author-Year	Unit of Analysis	Method	Main Output Variable
Present Paper	10-K: Risk factors in Item 1A	A new multi-label text classifier	25 types of risk factors.
Li (2008)	10-K: words	Content Analysis	Readability. No label.
Li (2010)	10-K: Sentences in Item 7 MD&A	Single-label classifier	3 labels
Feldman et al. (2009)	10-K: Words in Item 7 MD&A	Word categorization	Word counts in 2 labels
Balakrishnan et al. (2010)	10-K: MD&A Section	Single-label classifier	3 labels
Hanley and Hoberg (2010)	Sections in IPO prospectuses	Single-label classifier	Cosine-similarity
Antweiler and Frank (2004)	Postings on Yahoo! Finance	Single-label classifier	3 labels
Das and Chen (2007)	Postings on Yahoo! Finance	Voting of single-label classifiers	3 labels
Tetlock (2007)	Finance news articles	Word categorization	A sentiment measure by positive and negative word counts
Tetlock et al. (2008)	Finance news articles	Word categorization	A sentiment measure by positive and negative word counts
Loughran and McDonald (2009)	10-K	Word categorization	A sentiment measure by positive and negative word counts

2.3 Comparisons

The present study is different from the abovementioned articles in the following aspects. First, the unit of analysis in this study is "individual risk factor", which hasn't been investigated in the literature. In the existing literature, the unit of analysis includes using whole news article, whole 10-K or other SEC forms, one section (mostly the MD&A section) of 10-K, one sentence in MD&A section, or one posting on the Internet forum. Using "risk factors" as the unit for text classification is an important idea because individual risk factor has different meanings and by nature, each bullet of risk factors could convey more sophisticated information than the sentiment of a whole section reported in 10-K. Second, this study classifies each risk factor into 25 types to represent

the meaning of each risk factor. In sharp contrast, existing literature typically classifies its sample text into three types: positive, negative, and neutral. The proposed algorithm can achieve quite high accuracy even when classifying 25 types. From the information extraction's perspective, the present paper presents an interesting example that has the simplicity of text mining and at the same time the classified output could be almost as informative as those extracted by Natural Language Processing applications. The reason is that each type of risk factors itself conveys more sophisticated information than positive or negative labels on a paragraphs of text. Risk factors are unique in that keywords are typically representative enough for each type of risk factors, contributing to the high performance of text classifiers. Third, from the text classification's view, this is a multi-label classification problem and the present study proposes a new algorithm that works well in this context. Existing studies in accounting or finance either use a mathematical function of words counts or use existing single-label text classification algorithms as the main tool during the computational text analysis stage.

2.4 Finance or accounting data mining applications in Information Systems

There exists sparse literature in top information systems journals that apply data mining to examine finance or accounting issues. Sarkar and Sriram (2001) introduce an automated system based on Bayes classifier and decision tree classifier by using financial ratios as predictors of a bank's performance and the posterior probability of a bank's financial health. Baesens et al. (2003) present the results from analyzing three real-life credit-risk data sets using neural network rule extraction techniques. Huang et al. (2004a) apply support vector machines (SVM) to analyze the corporate credit rating analysis in the United States and Taiwan markets. Gu et al. (2007) examine how users value virtual communities and how virtual communities differ in their value propositions. In this study, a key variable is "posting quality" of messages in stock discussion boards. Messages are

classified by text classification algorithms into three categories: signal, noise, and neutral for hypothesis testing about the properties of the competition among virtual communities. In a series of papers, Schumaker and Chen (2009) develop a stock trading system by using quantitative portfolio selection strategies and news articles quantified by support vector machine regression and rigorous text mining techniques. The proposed system is shown to produce superior stock trading return in a short period of time. Cecchini et al. (2010) use support vector machine and basic financial data to predict fraudulent public companies cases. Bai et al. (2010) studied the data quality risk in accounting information systems.

3. DATA COLLECTION

3.1 SEC Form 10-K and Risk Factors

Risk factors are collected from all publicly-listed USA companies' SEC Form 10-K filings, which are publicly available from The Electronic Data Gathering, Analysis and Retrieval (EDGAR) database on the Internet. A Java program is written to download and to parse all 10-K files available from EDGAR.

In this study, 6,208 firms' 21,077 10-K files from 2006/1/1 to 2010/5/31 are collected in HTML format from EDGAR. The starting year is 2006 because since 2006, companies are required to report the "risk factors" in a separated section (Item 1A – Risk Factors) in their annual reports. According to Item 503(c) of Regulation S-K (§229.503(c) of this chapter), "Risk Factors" are defined as

> *Where appropriate, provide under the caption "Risk Factors" a discussion of the most significant factors that make the offering speculative or risky. This discussion must be concise and organized logically. Do not present risks that could apply to any issuer or any offering. Explain how the risk affects the*

issuer or the securities being offered. Set forth each risk factor under a subcaption that adequately describes the risk. The risk factor discussion must immediately follow the summary section.

For example, the headings (subcaption) of the first five risk factors in the Oracle's 10-K in 2010 are:

1. *"Economic, political and market conditions, including the recent recession and global economic crisis, can adversely affect our business, results of operations and financial condition, including our revenue growth and profitability, which in turn could adversely affect our stock price."*
2. *"We may fail to achieve our financial forecasts due to inaccurate sales forecasts or other factors."*
3. *"We may not achieve our financial forecasts with respect to our acquisition of Sun or our entrance into a new hardware systems business, or the achievement of such forecasts may take longer than expected. Our profitability could decline if we do not manage the risks associated with our acquisition and integration of Sun."*
4. *"Our success depends upon our ability to develop new products and services, integrate acquired products and services and enhance our existing products and services."*
5. *"Our strategy of transitioning from Sun's indirect sales model to our mixed direct and indirect sales model may not succeed and could result in lower hardware systems revenues or profits. Disruptions to our software indirect sales channel could affect our future operating results."*

These risk factors are reported as a list of bullets with detailed explanations (which are skipped for brevity here) right after each bullet. In this study, text classification is

applied to the headings but not to the detailed explanations because in our pilot study, text classification on both headings and descriptions lead to worse performance.[2]

Most companies report 15 to 50 risk factors in each annual report. For the same company, risk factors are quite similar across years. Typically, companies prepare 10-Ks by adding 1 to 5 risk factors to the previous year's 10-K after reordering some risk factors. This study also tracks the orders of appearance of risk factors in each 10-K and the information content embedded of orderings will be examined as well.

One limitation of this study is about using computer programs to parse and separate risk factor bullets in HTML files. The problem is that 10-Ks are submitted to SEC in non-standardized HTML files. Without the regulation change in 2005, it is quite difficult and almost impossible to locate and extract all risk factors by computer programs. Even after 2006 when risk factors were required to be summarized in Section Item 1A, it remains a challenging task to extract all risk factors because companies use very different HTML formatting templates to represent the beginning of the target section, "Item 1A Risk Factors". At the same time, extracting each bullet of risk factors and separating headings of risk factors from other parts are even more challenging. First, companies may use all kinds of HTML languages to represent a bullet point of risk factor. Specifically, companies may use different types of bullet tags, different types of numbering tags, various HTML templates, bold fonts, italic fonts, or a different font to represent the heading of a risk factor. Those headings are visually easy to be identified by human readers but difficult for computational parser to recognize. Second, companies may insert sub-section headings to represent a group of similar risk factors. For example, firms may insert "risks related to our products" before 3-5 risk factors, adding another layer of complexity to the parsing program. Last, few companies even violated the accounting

[2] The reason could be that detailed explanations include much more words than headings. Therefore, a keyword become more probable to appear in incorrect types' explanations, reducing the power of that keyword to identify a specific type of risk factors.

rules to report risk factors in other sections. For example, several companies "forgot" to adapt to new rules in 2006. Citibank always file 10-K in a different organization of sections. As a result, in the end, the parsing program can successfully extract roughly 75% of all 10-K forms and can collect 500,051 risk factors during the sample period.

3.2 Training Set and the Risk Factors Labeling

As a first step of implementing a supervised classification algorithm, researchers read hundreds of annual reports and subjectively identify 25 common types of risk factors. These risk factors types are explained in detail in Table 7 with examples of risk factors. The criteria of choosing these 25 types include the following. First, this list of definitions is expected to cover most frequently mentioned risk factors. In other words, this list is expected to be extensive but is not exhaustive because in practice, companies may report any type of risk factors in free forms. Some risk factors are very firm-specific risk factors and similar risk factors only appear in very few 10-Ks. Conceptually and in theory, there may exist hundreds or even thousands types of risk factors and this study focuses on a small set of common risk factors. Second, this list of risk factors is expected to be mutually exclusive by its meaning. Third, some less-frequently mentioned risk factor types are included because those risk factors could be influential. For example, Type 2, "restructuring risks", is expected to be less common across all companies but it could be very impactful. Clearly, this part involves the researcher's subjective judgment: some important risk factor types may be left out in this study.

Next, four student researchers in information systems are recruited to label 10,000 risk factors in fiscal year 2007. These 10,000 risk factors are chosen from 800 largest companies in 2007 in terms of total asset value. The researchers also strive to diversify the sample firms across industries so that the training set of risk factors won't be too similar by nature. Students have been briefed the definitions of risk factor types and they

are given a small number of examples of labeled risk factors to learn more about the correct classification of risk factors. Each student labeled 5,000 risk factors and each risk factor is labeled by two students. 2007 is chosen simply because it is the middle year of the sample period. Training set is built by one year's data because firms may report identical risk factors in different years: in the extreme case, firms may report an identical risk factor in 5 years in a row. The final training set is formed by using the records that two students completely agree on all types and also that risk factor is classified to at least one of the 25 types. The final training set includes 3,153 risk factors from 4,267 companies' 92,993 risk factors in 2007. In other words, we only know the true types of only 3,153 risk factors out of 500,051 risk factors collected in 5 years. 496,898 risk factors are treated as the test set and will be automatically labeled by the main classifier introduced in the next section.

As another example of risk factors labeling, the first five risk factors of Oracle mentioned in the preceding section are labeled as Type 8, Type 22, Type 4, Type 24, and Type 25, respectively. In the Oracle's examples, all five risk factors are labeled as single type. In general, one risk factor could be labeled as multiple types.

4. A MULTI-LABEL TEXT CLASSIFICATION ALGORITHM

This section first introduces a new algorithm and next evaluates its performance. The advantages of the proposed algorithm include: (1) its simplicity in terms of ease of understanding and implementation; (2) transparency of interpreting the results; (3) linear running time; (4) zero training time; (5) high scalability.

4.1 The algorithm

Step 1: TF-IDF (Han and Kamber 2006) is used to create a customized word vector of each risk factor.[3] The final word vector includes 1,430 composite words of one or two keywords. This word vector is chosen based on its performance in the training set. In the following paragraphs, X_i is used to denote the word vector of each unlabelled risk factors while T_k^j is used to denote the word vector in the training set with label type-*j*. The subscripts *i and k* are used to denote risk factor samples.

Step 2: For each unlabeled testing record, the cosine similarity (Han and Kamber 2006) with each record in the training set is calculated. SIM(X_i, T_k^j) is used to denote the cosine similarity.

Step 3: Next, 25 binary classifiers are used to classify records into each one of the 25 types. That is, the proposed algorithm will classify each record as 0 or 1 type-by-type 25 times. For type *j*, I compute the average of the top-N similarity from all training records in that type. This similarity value is defined as the similarity of the unlabelled testing record with that type of risk factors in the training set. For ease of exposition, I assume T_k^j be the record with the k^{th} largest similarity in type *j*. For a Top-*N* similarity algorithm, the similarity of X_i with type *j* is given by

$$\sum_{k=1}^{N} \mathrm{SIM}(\mathrm{X_i}, \mathrm{T_k^j})/\mathrm{N}$$

Step 4: For each type, if the Top-*N* similarity is greater than a threshold *K*, record *i* is labeled as 1 for type *j* and labeled as 0 otherwise.

[3] This study uses RapidMiner, an open source data mining tool based on Weka, to create the customized word vector. RapidMiner's built-in filters, include "English Stop Word List Filter", "Token Length Filter (>= 4)", "Porter Stemmer", and "Terms-2-Gram Generator", are used to create a candidate word vector. The word vector produced by RapidMiner contains 16,353 composite words. Next, words with low occurrences (low TF) or low informative value are dropped. Specifically, I drop all words with less than three occurrences in the training set. For words less than ten occurrences in the training set, I only keep the words with more than 75% conditional frequency in one type/category. In other words, only keywords with high predictive power are kept: i.e., when it shows up, with more than 75% the target risk factor belongs to a specific category in the training set.

As a result, there are two parameters in this algorithm: (1) using top-N records to represent the category's similarity, and (2) a threshold K for determining whether a testing record should be labeled as that type of risk factor. In general, different threshold values can be used for different types. But, this study keeps it simple to use the same threshold, K, for all types. The performance of a multi-threshold classifier is left for future research.

The running time of this algorithm is given by:

$$N_{Test} \times (T_T \times N_{Training} + T_S)$$

N_{Test}: the number of records in the test set;

T_T: time for computing similarity;

$N_{Training}$: the number of records in the training set;

T_S: time for sorting similarity.

The proposed algorithm's running time is linear in the size of the test set. The total labeling time of this algorithm is relatively long, compared with other data mining algorithms. However, this algorithm can be split into sub-tasks and computed parallely type-by-type plus record-by-record independently.

4.2 Performance Evaluation on the Training Set

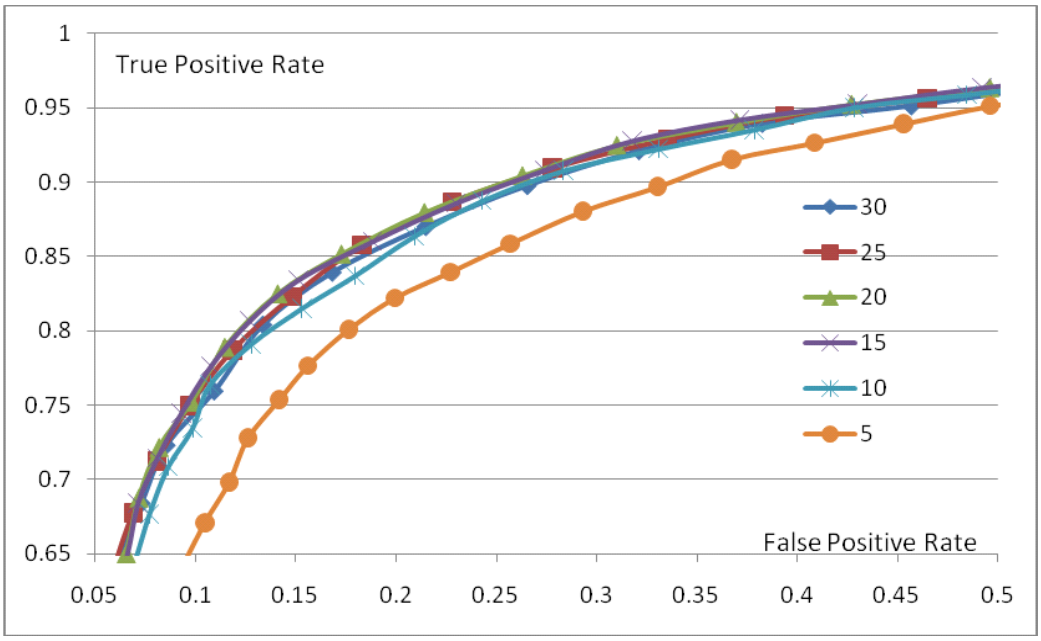
10-fold cross-validation and ROC curve (Provost and Fawcett 2001) are utilized to evaluate the performance of the proposed algorithm. Figure 1 reports the result of the ROC curve. Y-axis is the true positive rate whereas X-axis is the false positive rate. The ROC curve is created by varying threshold *K* from 0.01 to 0.50 with a step 0.01. Figure 1 reports six cases of *N* with values ranging from 5 to 30. We can observe that the performance of *N*=5 and *N*=10 are much worse than the other four classifiers. Among the

other four classifiers, *N*=15 and *N*=20 slightly outperforms the other two. The original intuition of using Top-N similarity is: the testing record may be very similar to a small subset of the training set, but very dissimilar to others in the same type in the training set. However, this experimental result only partly confirms this conjecture. It supports this view because there exists an optimal intermediate value of *N* that optimizes the performance of the proposed algorithm. However, it does not fully support this intuition because when *N* is smaller than 10, the performance of the classifiers are very poor, compared with the performance of classifiers with *N* greater than 10 whereas when *N* is larger than 10, the performance are similar.

Table 8 in the appendix provides more detailed comparison between *N*=15 and *N*=20, the two candidates of classifiers for our next-stage's study. The performance of classifiers is very close. *N*=15 and *K*=0.25 are chosen as the parameters of the main classifier. Table 9 reports the false positive rates and the conditional true positive rates of this main classifier. The main classifier will be applied to all unlabeled risk factors. Next section will examine the information content of these risk factors.

The high performance of the classifiers is due to the following reasons. Risk factors headings are typically concise, informative in roughly 20 words. These headings are relatively standardized across companies and years. Different companies may have exactly the same risk factor headings, possibly prepared by the same accounting firm or because the accountants-in-charge copy and paste from other firm's 10-K filings. Also, "keywords" are good representative of the meaning of those headings in most cases. For example, "intellectual property" along is quite informative in identifying risk factor types. In the "Risk Factors" section, when "intellectual property" appears, it is almost sure that risk factor is about intellectual property risks and also those two words rarely appear in other risk factor headings.

Figure 1 ROC Curves of the Proposed Algorithm



5. The Information Content of Risk Factors

This section uses conventional regression method in economics to examine the additional explanatory power of the 25 risk factors as independent variables to regress on ten different financial measures. Intuitively, risk factors should be highly correlated with at least risk measures.

5.1 Data Description

Dependent variables:

This study examines ten dependent variables that represent different aspects of firm risks and performance. Specifically, this study includes three measures of risks, two stock returns, and five financial ratios. All variables can be calculated by the public financial statement information available in Compustat from Wharton Research Data Services (WRDS). The definitions are given as follows:

(1) Implied Volatility calculated from Stock Options

In finance, the standard definition of the total risk of a company is the variance of stock returns of the target company. One estimate of this construct is the historical, realized variance of stock returns. Instead, this paper uses the implied volatility from stock options as the main measure of the target company's risk. The implied volatility is the volatility implied by the market price of the option based on Black-Scholes option pricing model Hull (2008). In other words, it is the volatility that yields a theoretical value for the option equal to the current market price of that option. Implied volatility is a better measure than the historical variance in this study because implied volatility is a forward-looking measure whereas historical volatility is an ex-post measure. For example, during the rise and down of oil prices in 2007 and financial crisis in 2008 in our sample, historical volatility is high but the expected ex ante volatility is samller.

(2) Debt Ratio. It is defined as (Total Liabilities)/(Total Assets).

Debt ratio indicates what proportion of debt a company has relative to its assets. It could be the simplest, most common financial ratio that is used to gauge the risk of a company, especially the risk to debt holders (Bharadwaj 2000). This measure gives an idea to the leverage of the company along with the potential risks the company faces in terms of its debt-load. The higher the ratio, the greater bankruptcy risk and also the greater risk will be associated with the firm's stock prices as well. A debt ratio of greater than 1 indicates that a company has more debt than assets, an important indicator that the target firm is on the edge or restructuring or bankruptcy.

(3-5) Annual Lowest Stock Return, Annual Stock Return, and Annual Highest Stock Return

Three stock returns are used as the dependent variable in the current study. These returns are calculated based on the closing stock of each fiscal year's end. Hence, the annual stock return is defined as the closing price of the next fiscal year divided by the closing price of the current fiscal year. The other two returns could be consider as upside or

downside stock risks. Annual lowest (highest) stock return is calculated by the lowest (highest) price during the following fiscal year divided by the closing price of the current fiscal year.

(6) Return on Assets (ROA). It is defined as (Net Income)/(Total Asset).

The return on assets (ROA) percentage shows how profitable a company's assets are in generating net income. ROA is one of the most common measures for studying firm performance in the business value of IT literature (e.g., Bharadwaj et al. (1999), Bharadwaj (2000), and Barua et al. (2004)).

(7) Gross Margin. It is defined as (Revenue - Cost of goods sold)/Revenue.

Gross margin is a common financial measure of the value proposition of the target firm (Barua et al. 2004). It shows how well a company manages its variable cost and how much value the buyers are willing to pay for the output of the target company.

(8) Net Margin. It is defined as (Net Income)/Revenue.

Net margin, or called profit margin, is a common financial ratio that shows the bottom-line of the target company. Compared with gross margin, net margin includes all other sources of costs and income and is a well-accepted measure of profitability of the target company.

(9) Tobin's Q. It is defined as (Market Capitalization + Total Liabilities)/(Total Assets).

In the business value of IT literature, Tobin's Q is also frequently used as an important performance measure (Bharadwaj et al. 1999). The assumption is that the market capitalization represents the "true value" of a company, including intangible assets (values) such as brand names and growth potential, which does not appear in other financial ratios.

(10) R&D Ratio. It is defined as (R&D expense)/Revenue.

This study also investigates this measure as a dependent variable because several risk factors types (types 10, 11, 16, and 24) seem to relate to R&D investments.

Independent variables:

Bear in mind that in Section 4, the unit of analysis is "individual risk factor" in text classification. In this section, the unit of analysis is a firm-year record represents each 10-K report in a yearly panel data. In other words, the original, untransformed "independent variable" is a matrix that represents all risk factors in each 10-K whereas now we need a method to condense that matrix into a row vector. In that matrix, each row represents a risk factor with positions represent the ordering of risk factors: i.e., the first row represents the first risk factor and the second row represents the second risk factors. Each column represents one of the 25 types of risk factors. As a result, we need a method to transform this matrix into a row vector for panel regression.

In the baseline case, this study uses 25 dummy variables to represent each 10-K in the panel regression. Each 0-1 dummy variable (binary variable) represents one type of risk factors. It is one as long as the associated type of risk factor is mentioned in any position in that 10-K whereas it is zero if that type of risk factor is never mentioned in the associated 10-K. In this baseline case, the ordering of risk factors is not used at all. Therefore, to utilize the risk factor ordering information, this study also creates two other sets of dummy variables for comparisons.

In the second case, only the first ten risk factors are used to calculate the 25 dummy variables. That is, a dummy variable is zero when that type of risk factor is never mentioned in that first ten risk factors in the target 10-K and it is one otherwise.

In the third case, this study considers only the first risk factor in each 10-K. A dummy variable is zero when the first risk factor does not belong to that type and it is one otherwise.

The frequency of each type of risk factors is summarized in Table 2. The last column reports the percentage of 10-Ks that contain at least one risk factor that is classified as a particular type. Some risk factors types, such as Type 5 (regulations change) and Type 20 (industry is competitive), are very common in 10-Ks whereas some types, such as Type 2 (restructuring risks) and Type 13 (infrastructure risks), are less mentioned in 10-Ks. One caveat is that these percentages are calculated not by true labels, but by the labels created by the text mining classifier, the accuracy of which is expected to be around 80%.

The 4th column, "First", reports the frequency of the first risk factor being classified as a particular type. Not surprisingly, the frequency drops sharply. Intuitively, the first risk factor should be the most unique, significant risk faced by the target firm and deserves special attention. As a benchmarking case, the first ten risk factors case is also analyzed. In the following analysis, the same regression analysis will be repeated three times for these three samples of independent variables.

Control variables:

This study includes two control variables. First, year dummy variables are used to control the yearly time effects. The reason is that between 2006 and 2010, the risk and performance of firms differ a lot because of the macroeconomic factors in each year. The second control variable is "firm size", which is an important factor that may affect the risk and financial ratios. Firm size in this study is defined as the logarithm of the number of employees obtained from Compustat.[4] Our main regression method is first-differencing panel regression. Therefore, one-year lagged dependent variable could be considered as a control variable for other firm-specific idiosyncratic errors.

[4] The logarithm of total asset is another common measure of firm size used in the literature. This study chooses the number of employees because total asset is used as the denominator in many of financial ratios of the dependent variables. In fact, using the logarithm of total asset leads to more significant results in the current study, probably because of larger sample size due to fewer missing values in "total assets".

Descriptive statistics of independent, dependent, and control variables are provided in Tables 2 and 3. In this sample, outliers are identified to be very influential. For example, when a company is at the edge of bankruptcy, all financial ratios may become quite different from normal companies. A typical strategy in finance (Barber and Lyon 1996) and accounting (Watson 1990) literature is Winsorization: to set all outliers to a specified percentile of the data rather than dropping records.

The current study uses a 95% Winsorization that sets all data with values below the 2.5% percentile at the values of the 2.5% percentile, and data with values above the 97.5% percentile at the values of 97.5% percentile. By this approach, outliers are not dropped but the influences of outliers are reduced. This approach is beneficial because outliers are important sample financial ratios for the present study. Those outliers are typically the companies with significant changes in operations, either bankruptcy or after merger and acquisition, both of which are of great interests for prediction in practice.

Table 2 Percentage of 10-Ks Contains Each Type of Risk Factors

Type	Name	Obs	First	Ten	All
1	Poor financial conditions	15731	7.42%	24.12%	38.05%
2	Undergoing restructuring	15731	0.31%	3.57%	6.24%
3	Inability to raise capital	15731	1.92%	14.01%	26.28%
4	Merger & Acquisition	15731	4.34%	35.48%	62.72%
5	Regulation or accounting rules change	15731	5.56%	56.40%	91.31%
6	Catastrophies	15731	1.10%	12.71%	33.72%
7	Shareholder's interest	15731	1.53%	9.67%	38.82%
8	Marcoeconomy risks	15731	12.96%	38.40%	54.06%
9	International risks	15731	0.85%	19.01%	42.43%
10	Intellectual property risks	15731	0.36%	14.89%	45.97%
11	Potential defects in products	15731	0.49%	13.57%	41.85%
12	Ongoing lawsuits	15731	0.74%	10.15%	28.77%
13	Infrastructure risks	15731	0.13%	2.54%	7.58%
14	Disruption of operations	15731	0.37%	10.84%	28.63%
15	HR risks	15731	1.75%	28.75%	68.98%
16	Licensing risks	15731	0.33%	5.50%	13.81%
17	Suppliers risks	15731	1.39%	27.88%	48.21%
18	Input price risks	15731	2.52%	18.95%	28.25%
19	Dependent on few large customers	15731	2.85%	21.43%	29.26%
20	Industry is competitive	15731	8.19%	57.81%	81.18%
21	Industry is cyclical	15731	3.50%	10.93%	14.31%
22	Volatile demand or financial results	15731	4.21%	24.80%	45.11%
23	Volatile stock price	15731	3.99%	18.09%	57.54%
24	RD & New product introduction risks	15731	2.26%	21.22%	29.96%
25	Downstream risks	15731	0.47%	7.97%	15.05%

Table 3 Summary Statistics of Key Variables (Winsorized at 95%)

Variable	Obs	Mean	Std. Dev.	Min	Max
Risk Measures					
Implied Volatility	11009	49.80%	22.07%	20.00%	100.00%
Debt Ratio	15731	57.47%	27.42%	12.24%	108.13%
Annual Lowest Stock Return	10331	-43.01%	25.91%	-89.00%	-3.40%
Firm Performance					
Return on Asset (ROA)	15731	-3.81%	19.08%	-62.68%	16.54%
Gross Margin	15731	36.01%	28.05%	-38.72%	81.17%
Net Income Margin	15731	-14.66%	57.66%	-224.72%	28.19%
Tobin's Q	15701	1.6589	1.0258	0.6548	4.5762
R&D Expense Ratio	6200	50.19%	114.94%	0.35%	469.16%
Return					
Annual Stock Return	10343	-5.46%	53.42%	-80.00%	127.27%
Annual Highest Stock Return	10311	46.42%	57.75%	0.84%	225.00%
Firm Size					
Total Asset (log)	15731	6.36	1.91	2.88	9.72
Number of Employees (log)	15078	0.01	1.92	-3.38	3.47

5.2 Regression Model

The baseline regression model is a first-differencing, firm-level panel regression.

$$Y_i^{t+1} - Y_i^t = \beta_0 + \sum_{t=2005}^{2010} \beta_t D_t + \sum_{j=1}^{25} \beta_j X_{ij}^t + \beta_s lnS_i^t + \varepsilon_{it}, \qquad (1)$$

where the left-hand-side (LHS) is the difference between the current financial ratio and the next year's financial ratio for prediction, D_t is the year dummy variable that controls

the time effects, X_{ij}^{t} represents the independent variables (25 dummy variables for risk factor types) for firm i, and S_i^t represents the firm size.

This study chooses the first-differencing model over fixed-effect panel regression following the exact recommendation by Wooldridge (2002) and Wooldridge (2009). The reason is that first-differencing has the advantage of being straightforward in virtually any econometrics package, and it is easy to compute heteroskedasticity-robust statistics in the first-differencing regression. Both fixed-effect and first-differencing panel regressions are unbiased and consistent for linear regression models. This study's dataset has a large number of firms but small number of years with serial correlated error terms, first-differencing is more efficient than fixed-effect model when being applied to this kind of sample (Wooldridge 2009). Serial correlation is typically high on financial time series data because firm performance improves/deteriorates based on previous year's performance. As a result, first-differencing is chosen as the regression model. This study does not choose random effect panel regression because in all regressions, random effect models are rejected at 99% significance level by Hausman's specification test.

The baseline regression (1) is first estimated by the standard least square method. Breusch-Pagan test for heteroskedasticity suggests the existence of heteroskedasticity at 99% level. As a consequence, the baseline regression will be estimated with a robust variance-covariance matrix in all cases in the following analysis (Wooldridge 2009).

Detailed estimation results for 10 dependent variables by using the first, the first ten, and all risk factors are reported from Table 10 to Table 15 in the appendix. The purpose of the current section is not to build a prediction model for any one of the ten dependent variables but to show that the extracted risk factors contain information content that could potentially contribute to building a better prediction model. The reason is that to rigorously predict stock returns, volatility, or earnings, we need to consider a more advanced econometrics model specialized in that prediction purpose. In finance or

accounting, a typical research objective of one academic paper is to predict one or two of the ten variables and therefore it is clearly too ambitious for the present paper to claim that risk factors can improve the prediction of all ten variables.

The goal of this paper is to show that extracted risk factors indeed correlate with those financial numbers and do contain abundant information for building business intelligence applications or future research opportunities. The next section will provide evidence to support this claim.

5.3 Empirical results and discussions

Proposition 1: *Risk factors can provide additional explanatory power for risk, stock return, and financial ratios.*

Proposition 1 is tested by comparing the R^2 from estimating (1) with the R^2 from estimating the regressions with only control variables. Formally, the benchmarking case is given by

$$Y_i^{t+1} - Y_i^t = \beta_0 + \sum_{t=2005}^{2010} \beta_t D_t + \beta_s lnS_i^t + \varepsilon_{it}, \qquad (2)$$

Intuitively, the difference in R^2 represents the additional explanatory power provided by the 25 risk factors types. The statistical testing method is the standard F-test. Results are reported in Table 4.

The last two columns in Table 4 report the baseline results, which suggest that including 25 dummy variables provides additional explanation power at 99% significance level for almost all dependent variables. When we use the first ten risk factors, the explanatory power of risk factors is still very significant at 99% significance level for all ten dependent variables. Lastly, when using only the first risk factor for prediction, only 5 out of 10 dependent variables show significant results. These three cases clearly show that risk factors contain information that is statistically significant.

Table 4 F-Test of the Information Content of 25 Risk Factors

	First Risk Factor		First Ten Factors		All Factors	
	F	p	F	p	F	p
Implied Volatility	3.43	0.0000	5.77	0.0000	6.67	0.0000
Debt Ratio	1.18	0.2478	2.74	0.0000	2.42	0.0001
Annual Lowest Return	1.50	0.0511	2.14	0.0008	2.58	0.0000
Annual Stock Return	1.18	0.2409	1.85	0.0061	1.48	0.0579
Annual Highest Return	1.33	0.1249	1.88	0.0051	1.63	0.0251
Return on Asset (ROA)	1.08	0.3620	2.25	0.0003	2.02	0.0019
Gross Margin	2.74	0.0000	5.27	0.0000	5.65	0.0000
Net Income Margin	2.11	0.0010	3.30	0.0000	4.11	0.0000
Tobin's Q	1.34	0.1187	3.15	0.0000	3.99	0.0000
R&D Expense Ratio	1.51	0.0536	1.94	0.0034	3.12	0.0000

These three results also suggest that using more risk factors does not seem to dilute the informational content of risk factors. The alternative possibility is “only the first risk factor or the first several risk factors contribute to explanatory power”. Table 4 provides preliminary evidence that shows using more risk factors produces higher overall explanatory power. The next step is to examine the information content of risk factors position-by-position to confirm this observation.

Next, this study investigates the information content of the risk factor *orderings*. Intuitively, there are three possibilities: (1) firms report risk factors in a completely random orderings. Any types may appear in any position with the same probability. (2) the risk factor ordering is not randomly generated but it does not provide additional information content. There are two possible explanations. First, accounting managers may put “industry is competitive” in an earlier order because “competition” is generally regarded as an important factor when evaluating risk or profitability. However, it does not mean that risk factor is more important to the target company. The other possibility is the “herding phenomenon”: accounting managers report a factor earlier in 10-K because all other 10-K follows similar patterns. (3) Firms report risk factors by following the

importance of risk factors. The target company will report the risk factors that they feel most influential in the earlier position. If this is true, risk factors appear in earlier positions will provide higher explanatory power than the risk factors reported later in the list. This study shows that the third possibility is true.

Proposition 2a: *Risk factors are not reported in a random order.*

Table 4 shows that the probability of occurrences in the first risk factor is different from the other 9 positions in the first ten risk factors. The idea of the hypothesis testing is sketched but the detail is omitted for brevity.

To test whether type-1 risk factor shows up in any position with the same probability, we first assume the probability of type 1 in any position is 7.42% (column 2 in Table 4). By the Binomial distribution's formula, the probability that there is at least one type-1 risk factor in the first ten risk factors is 53.74%. Next, we can compare this value with 24.12% in column 3 of Table 4. by the Probability Ratio Test with a sample size 15731. Clearly, the equivalence hypothesis is rejected at 99% significance level in this case. As another example, to test type 4, the probability is assumed to be 4.34% and the probability that there is at least one type-4 in the first ten risk factors is 35.83%, which is very close 35.48% (column 3 of Table 4). As a result, we cannot reject the hypothesis that type 4 appears in any position with probability 4.34%. By this hypothesis testing method, most of the risk factors can be shown to appear in a non-random fashion.

Proposition 2b: *Risk factors in the earlier position provide larger explanatory power.*

To test this proposition, we examine the incremental explanation power of risk factors in each position. Specifically, we first estimate equation (1) by using only the X^{th} risk factor in 10-K. The R^2 from estimating (1) is compared with the R^2 from estimating (2), which includes only the control variables. Again, the standard F-test is used to

examine the additional explanatory power provided by the risk factors in X^{th} position. F-values from F-tests are visually depicted in Figure 2 and Figure 3. In Figure 2, the x-axis value means using the X^{th} risk factor for prediction. The y-axis is the F-value from the F-test that compares regressions with/without X^{th} risk factors. In other words, F-value in this figure is a measure of the explanatory power of X^{th} risk factor.

Figure 2 and especially Figure 3 suggest that the explanatory power of risk factors seems to have a U-shape in terms of its position. The first few risk factors indeed provide larger explanatory powers than factors in other positions, especially factors around 20^{th} position. The decline in explanatory power is not that obvious and also fluctuates a lot. The other observation is that the explanatory power for financial ratios (Figure 3) seems to decline more sharply than the dependent variables in Figure 2. Surprisingly, the explanatory power of the risk factors appear in the end of this list also shows larger explanatory power. But, one very important caveat is that different 10-K form reports different number of risk factors. Hence, a large number of position also implies larger total number of risk factors, which results only from a unique group of companies' 10-K. As a result, even the last few risk factors are very significant, it could due to those companies are very different from the companies that report 10 to 20 risk factors. Proposition 3 will examine this issue.

Selected cases are reported in Table 5. In all three cases, none of the individual risk factors contains significant explanatory power for the dependent variables, a sharp contrast to the first risk factor in Table 4. The 7^{th} risk factor is reported because each of the first 6^{th} risk factors provides significant explanatory power in at least one dependent variable at 95% significance level. In sharp contrast, there is no significant result in the three risk factors in Table 5 at even 90% significance level.

Figure 2 The F-Test Values of Using One Risk Factor to Predict Risk and Return measures

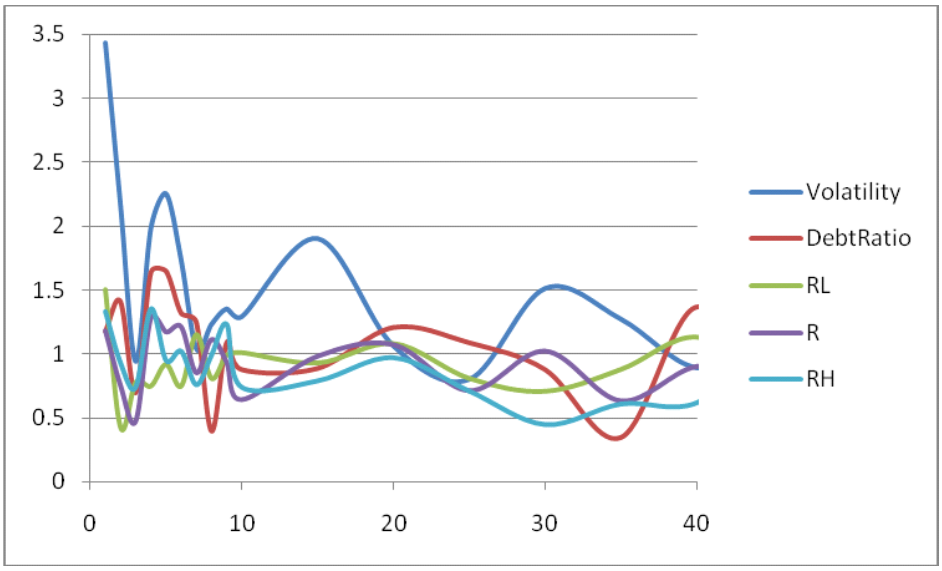
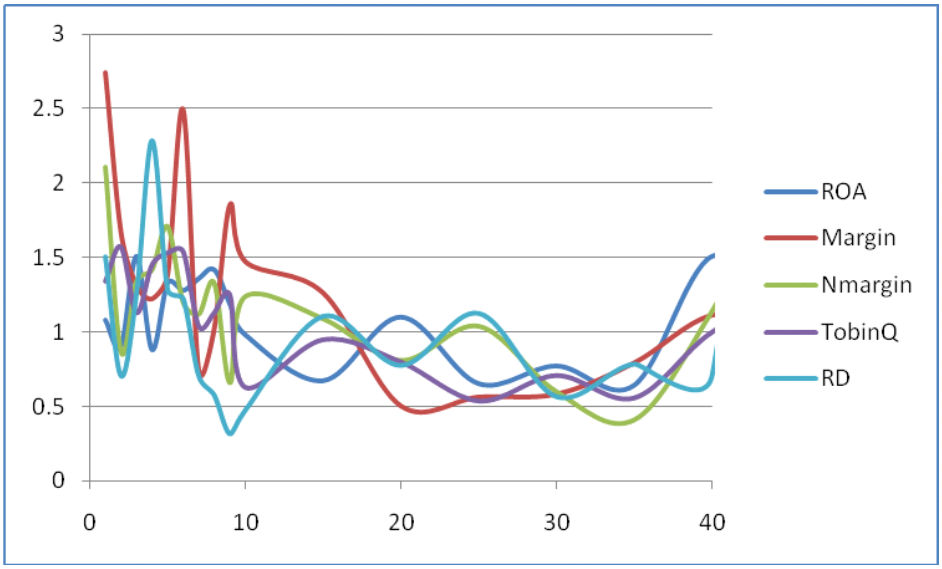


Figure 3 The F-Test Values of Using One Risk Factor to Predict Financial Ratios



The x-axis means the target risk factor is the x^{th} risk factor reported in 10-K.
The y-axis is the F-Value from the F-Test.

Table 5 F-Test of the Information Content of 7th, 15th, and 25th Risk Factors

	7th Risk Factor		15th Factors		25th Factors	
	F	p	F	p	F	p
Implied Volatility	1.04	0.4144	1.90	0.0043	0.80	0.7471
Debt Ratio	1.25	0.1783	0.89	0.6221	1.09	0.3448
Annual Lowest Return	1.15	0.2744	0.93	0.5640	0.81	0.7378
Annual Stock Return	0.85	0.6788	0.98	0.4968	0.71	0.8530
Annual Highest Return	0.76	0.7923	0.79	0.7624	0.71	0.8495
Return on Asset (ROA)	1.36	0.1066	0.67	0.8916	0.65	0.9087
Gross Margin	0.74	0.8258	1.26	0.1766	0.56	0.9596
Net Income Margin	1.12	0.3116	1.09	0.3444	1.04	0.4077
Tobin's Q	1.04	0.4048	0.95	0.5279	0.54	0.9688
R&D Expense Ratio	0.71	0.8493	1.11	0.3232	1.13	0.3030

Table 6 F-Test of the Information Content of the Number of Risk Factors

	First Factor		First Ten Factors		All Factors	
	F	p	F	p	F	p
Implied Volatility	0.01	0.9280	1.35	0.2450	13.14	0.0003
Debt Ratio	3.45	0.0634	0.52	0.4690	0.23	0.6279
Annual Lowest Return	4.80	0.0284	2.92	0.0874	7.08	0.0078
Annual Stock Return	23.28	0.0000	15.48	0.0001	11.05	0.0009
Annual Highest Return	22.44	0.0000	16.08	0.0001	10.71	0.0011
Return on Asset (ROA)	1.92	0.1662	1.74	0.1876	5.39	0.0203
Gross Margin	7.12	0.0076	3.25	0.0714	2.13	0.1447
Net Income Margin	8.51	0.0035	3.63	0.0567	3.01	0.0829
Tobin's Q	0.33	0.5645	1.28	0.2573	10.92	0.0010
R&D Expense Ratio	0.08	0.7756	0.30	0.5857	0.09	0.7684

Proposition 3: *The number of risk factors can provide explanatory power for risk, stock return, and financial ratios, in addition to the risk factor types.*

This proposition is tested by applying F-test to compare the R^2 from estimating (1) and the R^2 from estimating (1) with the number of risk factors as one more independent variable. Results are reported in Table 6. 18 out of 30 cases are significant at 10% level

while 11 out 30 cases are significant at 1% level. This provides strong evidence that the number of risk factors also contributes to additional explanatory power in addition to the risk factors types. There are several potential explanations behind this observation. First, more risk factors could mean the target company indeed face more types of risks and bear higher overall risks. This could explain the correlation to risk measures and stock returns (higher risks require higher stock return to compensate equity holders, a well accepted theory and fact in finance). Second, the number of risk factors could be a proxy variable of the quality of the 10-K form. Firms with limited resource typically prepare shorter list of risk factors in 10-K. As a result, 10-K could be a proxy variable of the quality of auditing or quality of back office operations of the target company, which may explain the correlation of the number of risk factors to some financial ratios.

5.4 Explanatory power of individual risk factor types

Regression results of our baseline cases are reported in the appendix from Table 10 to Table 15. This study uses the number of stars of significance in these six tables as a measure of the informative contents of each risk factor type. The total number of stars of significance is summarized in Table 16.

Type 1 and Type 8 are the two most informative risk factors, followed by Type 17, Type 19, and Type 22. In practice, Type 1 (poor financial condition) could be intuitively the most informative risk factor that analysts read carefully when they search for valuable information in the 10-K. Type 1 is also the most informative type in this analysis: it is particularly significant when regressing on five financial ratios. Type 8 (macroeconomic risks) is the second most informative risk factor type. The reason could be the current sample period covers financial crisis period. First, if managers truthfully report risk factors, firms with larger macroeconomic risk clearly ended up hurting the most during the sample period. Second, if managers report risk factors to fulfill their own interests,

executives knew in advance that their operating performance would be far below average during the financial crisis could blame the poor performance on the external macroeconomic risk to reduce their own responsibilities of poor management.

Type 23 (volatile stock price risks) is the least informative risk factor in this study. This is a surprising result because it should be at least correlated with the implied volatility. One explanation could be more than 50% of the 10-Ks contain this risk factor, which makes it too generic. The other explanation is this risk factor is correlated with several other risk factors, lowering its explanatory power. Except Type 23, there are several other risk factors with low explanatory power. This list includes Types 2, 11, 12, 14, 15, 16, 20. Some of these risk factors are intuitively uninformative even before the rigorous regression analysis. For example, Type 20, "operating in a competitive industry" is a highly mentioned risk factor type. In our sample, more than 80% of the companies mentioned this type of risk. Even the "seemingly monopoly" Microsoft claimed that they operate in a highly competitive industry. Another similar example is Type 15 (human resources risks), which appears in roughly 70% of 10-Ks. This risk factor generically states the importance of executives and R&D staffs. For example, Yahoo! stated the importance of Jerry Yang, their co-founder and CEO, in 2007's 10-K. Mr. Yang turned down the acquisition offer from Microsoft at $31 Jan 31 2008. Later, Mr Yang stepped down as the CEO in November 2008 when Yahoo! was traded around $10. Type 15 may only imply the importance or control power of their key staff but may not imply the "risks" of losing that key staff. In contrast, Apple never disclosed the health risks of its key staff and the risks of losing that key staff in 10-K.

Different from other empirical studies, these "insignificant findings" about uninformative risk factors also have important policy implications. According to the requirement of SEC stated in Section 3.1, firms should report the most significant factors and *"Do not present risks that could apply to any issuer or any offering."* However, this

study results suggest firms may not meet this requirement when preparing SEC Form 10-Ks. Accounting regulation authority should enforce this rule because reducing the uninformative risk factors in 10-K can improve the information disclosure quality of the 10-K, saving company's preparation costs and 10-K readers' time.

6. CONCLUSION

The most important contribution of the present study is to identify a valuable text classification application in financial accounting. This research also provides a new algorithm to quantify risk factors reported in 10-K forms. These risk factors could play an important role in building real world business intelligence systems or for more sophisticated academic research along several directions. First, risk factors can be used to evaluate the risks to corporate debt holders, such as banks or corporate bond investors. Specifically, risk factors can be incorporated into the existing corporate credit rating systems (such as S&P, Moody's, or Fitch) or the cost of capital models in accounting or finance studies. The second natural direction is to use risk factors to investigate the risks to equity holders. The present study reported preliminary results about the correlation between risk factors and stock volatility and stock returns. Future studies can use more elaborate volatility or stock return models in finance, such as Fama-French three factor models, to study the impacts of risk factors. Specifically, risk factors may contribute to market risk (systematic risk), industry level risk, or firm-specific risk. Different levels of risks may contribute to the determination of stock prices in different ways. Third, risk factors could be included in the existing earnings prediction models, such as the studies listed in the literature review, to help stock analysts, fund managers, or individual investors make better investment decisions. Compared to abundant marketing data mining studies and applications (e.g., Padmanabhan and Tuzhilin (2003), Huang et al. (2004b), Padmanabhan et al. (2006), Huang et al. (2007), Bai (2010), Wei et al. (2010)),

investment analytics is an underexplored area for information systems researchers. Forth, one important usage of 10-K is management fraud detection. There exists a long stream of literature that attempts to find indicators of potential management fraud events. For example, Cecchini et al. (2010) use support vector machines and basic financial data to predict fraudulent public companies cases. Risk factors could potentially contribute to this line of research. Last, monitoring risk factors of companies in the related industries may help managers better gauge upstream supply chain risks and downstream demand risks by building better business intelligence systems to improve supply chain management and demand forecasting tasks.

The present research has several limitations. In the text mining phase, this paper only uses a simple, new text mining algorithm to extract risk factors. Although the performance is satisfactory, there may exist better algorithms that significantly outperform the proposed algorithm. For example, support vector machine could be used in this context. Performance of classifiers could be improved also by hierarchical classification or by using advanced natural language processing techniques. The definition of risk factor types could be refined and more risk factors types could be classified. For example, this study defines risk factor types at a relatively high level. Within each type, there are three to five detailed sub-categories that could provide even more information.

I hope that this research demonstrates a first small step to automate extracting risk factors or other valuable textual information in SEC filings. This paper has shown that the proposed algorithm can extract risk factors at a satisfactory performance and from the empirical results, the extracted risk factor information seems to contain abundant information for practical use or for further academic studies.

Appendix

Table 7 Risk Factor Categorizations and Definitions

No	Name	Definitions and examples
1	Poor financial condition risks	Factors related to history of loss, resulting in poor financial conditions. e.g., "*We have experienced net losses, and we may not be profitable in the future.*"
2	Restructuring risks	The target company has filed bankruptcy protection, or the company mentioned it is undergoing restructuring. e.g., *"We may need to incur impairment and other restructuring charges, which could materially affect our results of operations and financial conditions."*
3	Funding risks	Inability to raise capital to expand, for normal operations, or match competition. e.g., *"Banctrust may need to raise capital in the future when capital may not be available on favorable terms or at all."*
4	Merger & Acquisition risks	Anything factors related to M&A: e.g., Acquisitions may not meet expectation, or the M&A cost is high. e.g., *"Implementing our acquisition strategy involves risks, and our failure to successfully implement this strategy could have a material adverse effect on our business."*
5	Regulation changes	Any risks about government regulation changes, including environmental, accounting, or privacy laws. e.g., *"The company is subject to environmental regulations and liabilities that could weaken operating results."*
6	Catastrophes	Natural disasters or terrorists attack. e.g., *"Future terrorist attacks may have a material adverse impact on our business."*
7	Shareholder's interest risks	This includes: (1) The holder's interest is different from the shareholders (2) the shareholder has very strong control power (few large shareholders) (3) no large shareholder. e.g., *"We may encounter conflicts of interest with our controlling stockholder"*
8	Macroeconomic risks	For example, economic downturn, financial crisis, high energy price, inflation, recession, or unemployment. e.g., *"Demand for our products will be affected by general economic conditions."*
9	International risks	Anything related to global operations, including currency/exchange rate risks. e.g., *"Our international operations are subject to many uncertainties, and a significant reduction in international sales of our products could adversely affect us."*
10	Intellectual property risks	This includes the target company may infringe or be infringed by other company's patents. e.g., "*we may not be successful in adequately protecting our intellectual property.*"
11	Potential defects in products	Product liabilities or any risks related to product defects. e.g., *"We may incur substantial costs as a result of warranty and product liability claims which could negatively affect our profitability."*
12	Potential/Ongoing Lawsuits	Current/ongoing significant litigation or lawsuits. e.g., *"We are currently subject to securities class action litigation, the unfavorable outcome of which might have a material adverse*

		effect on our financial condition, results of operations and cash flows."
13	Infrastructure risks	Risks related to changes, upgrades, maintain the target company's infrastructure, which includes distribution network, IT, or organizational infrastructure. e.g., *"The infrastructure of our transmission and distribution system may not operate as expected, and could require additional unplanned expense which would adversely affect our earnings."*
14	Disruption of operations	Risks about operations may be disrupted due to complex manufacturing process or software systems. e.g., *"Material disruption to our manufacturing plants in Wisconsin could adversely affect our ability to generate revenue."*
15	Human resource risks	Risks about attracting, recruiting, maintaining key personnel or employees, such as CEO, executives, R&D staff, or sales people. e.g., *"We depend upon our key personnel and they would be difficult to replace."*
16	Licensing related risks	Dependent on other company's technology licensing or government license to operate business. e.g., *"If we are unable to renew our licenses or otherwise lose our licensed rights, we may have to stop selling products or we may lose competitive advantage."*
17	Suppliers risks	Any risks related to upstream suppliers, including OEM manufacturers. e.g., *"A change in sales strategy by the company's suppliers could adversely affect the company's sales or earnings."*
18	Input prices risks	Any description about the input prices (raw material prices) may go up. e.g., *"Our inability to pass through increases in costs and expenses for raw materials and energy, on a timely basis or at all, could have a material adverse effect on the margins of our products."*
19	Concentration on few large customers	High concentration on few large customers. e.g., *"Our sales could be negatively impacted if one or more of our key customers substantially reduce orders for our products"*
20	Competition risks	Industry is competitive, strong competition, or increasing competition. e.g., *"We compete in distribution industries that are highly competitive and we may not be able to compete successfully."*
21	Industry is cyclical	Industry is cyclical. e.g., *"We operate in an industry that is cyclical and that has periodically experienced significant year-to-year fluctuations in demand for vehicles."*
22	Volatile demand and financial results	Demand and/or Financial results are volatile and unpredictable. e.g., *"Our future revenue, gross margins, operating results and net income are difficult to predict and may materially fluctuate."*
23	Volatile stock price risks	The target company's stock price is volatile. e.g., *"The price of our common stock has fluctuated widely in the past and may fluctuate widely in the future."*
24	New product introduction risks	Potential delays or fails in new product introduction, or the new production introduction is critical to the target company's success. e.g., *"Our success depends on our ability to successfully develop and commercialize additional pharmaceutical products"*
25	Downstream risks	Risks associated with distributors or retailers. e.g., "*We face a number of risks related to our product sales through intermediaries.*"

Table 8 Performance Comparison between N=15 and N=20

	N=15		N=20	
K	False Positive	True Positive	False Positive	True Positive
0.16	55.21%	97.08%	42.72%	95.21%
0.17	49.22%	96.35%	36.98%	94.01%
0.18	43.02%	95.24%	31.01%	92.45%
0.19	37.14%	94.20%	26.29%	90.42%
0.20	31.75%	92.77%	21.40%	87.95%
0.21	27.39%	90.80%	17.26%	85.16%
0.22	22.97%	88.58%	14.07%	82.49%
0.23	18.83%	86.01%	11.43%	78.91%
0.24	15.11%	83.41%	9.78%	75.23%
0.25	**12.67%**	**80.65%**	8.15%	72.19%
0.26	10.75%	77.64%	7.15%	68.79%
0.27	9.27%	74.47%	6.52%	65.05%
0.28	8.08%	71.46%	5.61%	61.34%
0.29	7.11%	68.41%	4.99%	56.74%
0.30	6.57%	64.95%	4.62%	52.36%

Table 9 A Multi-Label Classification Confusion Matrix for the Main Classifier with N=15 and K=0.25

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	
1	24	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	92%
2	0	34	1	5	1	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	77%
3	1	0	30	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	83%
4	2	0	0	242	5	0	1	0	5	1	0	2	0	0	0	0	1	0	0	2	0	0	4	0	0	91%
5	1	0	0	0	438	0	1	2	7	0	1	8	0	0	2	0	0	1	1	10	0	1	5	0	0	92%
6	0	0	0	0	1	92	0	1	2	0	0	1	0	5	0	0	0	0	1	0	0	0	0	0	0	89%
7	0	0	0	0	0	0	38	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	7	0	0	81%
8	0	0	0	1	5	1	0	86	7	0	0	0	0	0	0	0	0	1	0	2	7	0	1	0	0	77%
9	0	0	0	1	8	1	0	4	77	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	85%
10	0	0	0	0	1	0	0	0	0	220	1	2	0	0	1	4	4	0	0	1	0	0	0	2	0	93%
11	0	0	0	0	4	0	0	0	0	1	87	1	0	0	0	0	1	1	0	1	0	0	0	1	0	90%
12	0	0	0	1	13	0	0	0	1	4	2	66	0	0	0	0	0	0	0	0	0	0	1	0	0	75%
13	0	0	0	0	1	0	0	0	0	0	0	0	14	4	0	0	1	1	0	1	0	0	0	0	0	64%
14	0	0	0	2	2	5	0	0	0	0	0	0	1	44	0	0	2	1	1	0	0	0	1	0	0	75%
15	0	0	0	1	1	0	0	0	0	0	0	0	0	0	176	0	0	0	0	1	0	0	0	0	0	98%
16	1	0	0	0	6	0	1	0	0	4	0	0	0	0	0	20	4	0	2	0	0	0	0	0	0	53%
17	0	0	0	1	2	0	0	0	1	1	0	0	0	2	0	0	103	4	1	2	0	0	0	1	0	87%
18	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	4	60	0	3	0	2	4	0	0	79%
19	0	0	0	0	2	0	0	0	0	0	0	0	0	0	1	0	2	0	61	5	0	0	0	2	0	84%
20	0	0	0	2	5	0	0	1	0	0	0	0	0	0	0	0	1	1	1	323	2	2	0	5	0	94%
21	0	0	0	0	1	0	0	6	0	0	0	0	0	0	0	0	0	0	0	6	48	7	0	1	0	70%
22	0	0	0	1	4	1	0	2	1	0	0	1	0	0	0	0	0	0	1	2	0	56	9	0	0	72%
23	0	0	2	2	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	7	83	0	0	86%
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	2	0	0	5	0	0	0	91	0	90%
25	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	3	0	2	1	0	0	0	0	30	79%
	83%	100%	91%	93%	87%	92%	88%	83%	75%	95%	96%	80%	93%	79%	96%	83%	80%	86%	85%	88%	84%	75%	71%	88%	100%	

The first column represents the true type. The first row represents the predicted type. The number in each cell is the number of predicted types, which is different from the numbers in the standard single-label classification confusion matrix. The percentages are simply the conditional probability of each row and columns.

Table 10 Predicting Risk Measures and Stock Returns by the First Risk Factor

	(1) Option Volatility	(2) Debt Ratio	(3) Lowest Annual Stock Return	(4) Annual Stock Return	(5) Highest Annual Stock Return
Type 1	-0.0109**	0.00467	0.0127	0.0143	0.0397
	(0.032)	(0.473)	(0.440)	(0.721)	(0.393)
Type 2	0.0500**	0.0121	0.0373	-0.00128	-0.211
	(0.040)	(0.509)	(0.606)	(0.995)	(0.190)
Type 3	0.0396***	0.00779	-0.0230	0.0870	0.125
	(0.002)	(0.562)	(0.537)	(0.308)	(0.177)
Type 4	-0.00747	0.00202	0.0217	0.0669	0.0915**
	(0.154)	(0.712)	(0.229)	(0.104)	(0.027)
Type 5	-0.0102**	-0.00694*	0.00297	-0.0493	-0.0660*
	(0.041)	(0.065)	(0.856)	(0.167)	(0.071)
Type 6	-0.00455	-0.00654	-0.0277	-0.107*	-0.135**
	(0.580)	(0.306)	(0.272)	(0.064)	(0.011)
Type 7	0.0196*	-0.0138**	-0.0247	-0.0767	-0.0280
	(0.064)	(0.013)	(0.362)	(0.200)	(0.641)
Type 8	0.00151	-0.00146	-0.0290**	-0.0602**	-0.0418
	(0.672)	(0.591)	(0.017)	(0.027)	(0.126)
Type 9	-0.000162	-0.00941	-0.0389	-0.0653	-0.0731
	(0.985)	(0.306)	(0.268)	(0.511)	(0.411)
Type 10	-0.0302	0.0112	-0.0464	-0.114	-0.189
	(0.189)	(0.482)	(0.418)	(0.357)	(0.132)
Type 11	0.000291	-0.0189**	-0.0475	-0.0205	-0.0442
	(0.978)	(0.031)	(0.286)	(0.837)	(0.639)
Type 12	-0.0200*	-0.00536	0.00264	0.0108	-0.0176
	(0.071)	(0.604)	(0.945)	(0.917)	(0.851)
Type 13	0.0164	-0.0362***	-0.0960	-0.246***	-0.157*
	(0.236)	(0.000)	(0.205)	(0.002)	(0.050)
Type 14	-0.0247*	0.0247	0.0187	0.0325	0.0393
	(0.088)	(0.194)	(0.739)	(0.817)	(0.743)
Type 15	0.00412	0.0135	-0.00155	-0.0861	-0.0871
	(0.682)	(0.139)	(0.954)	(0.156)	(0.171)
Type 16	-0.0103	-0.00437	-0.0205	0.0245	-0.0226
	(0.555)	(0.655)	(0.693)	(0.791)	(0.819)
Type 17	-0.0403***	-0.0225**	-0.0262	-0.0747	-0.0716
	(0.000)	(0.029)	(0.402)	(0.299)	(0.301)
Type 18	0.0183***	0.00216	-0.0335	-0.0525	-0.0641
	(0.007)	(0.735)	(0.151)	(0.341)	(0.250)
Type 19	-0.0188***	-0.00135	-0.0354	-0.0488	-0.0667
	(0.004)	(0.832)	(0.113)	(0.352)	(0.211)
Type 20	-0.0127***	-0.000667	-0.00523	-0.0108	-0.0132
	(0.009)	(0.842)	(0.698)	(0.715)	(0.655)
Type 21	-0.0154**	-0.00180	-0.0473**	-0.0315	-0.0557
	(0.026)	(0.705)	(0.012)	(0.452)	(0.216)
Type 22	-0.0169***	0.00136	0.0365*	0.0465	0.0575
	(0.001)	(0.800)	(0.061)	(0.273)	(0.192)
Type 23	-0.00657	-0.00999*	0.0171	0.0347	0.0264
	(0.292)	(0.082)	(0.406)	(0.456)	(0.611)
Type 24	-0.00977	0.0178*	0.0395	0.0665	0.0459
	(0.259)	(0.058)	(0.137)	(0.279)	(0.516)
Type 25	0.0197	-0.00146	-0.0612	-0.130	-0.138
	(0.316)	(0.939)	(0.172)	(0.167)	(0.110)
log_emp	0.00196***	-0.00162**	-0.00454**	0.00196	-0.00565
	(0.003)	(0.013)	(0.018)	(0.658)	(0.243)
N	6559	10057	5958	5965	5929
R^2	0.184	0.0282	0.372	0.365	0.132

p-values in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Year Dummies and a constant are omitted.

Table 11 Predicting Financial Ratios by the First Risk Factor

	(1) ROA	(2) Gross Margin	(3) Net Margin	(4) Tobin's Q	(5) R&D Ratio
Type 1	0.0135*	0.0221***	0.0654***	-0.0293	-0.125***
	(0.053)	(0.005)	(0.001)	(0.393)	(0.005)
Type 2	-0.00698	-0.0218	-0.0489	0.114	-0.250
	(0.793)	(0.562)	(0.534)	(0.322)	(0.359)
Type 3	0.0117	0.0131	0.0512	0.0448	0.0444
	(0.447)	(0.381)	(0.249)	(0.472)	(0.623)
Type 4	0.00172	-0.0109*	-0.0105	-0.00453	-0.0265
	(0.790)	(0.072)	(0.510)	(0.857)	(0.610)
Type 5	0.0104**	0.00783	0.0163	0.0206	-0.00224
	(0.025)	(0.200)	(0.303)	(0.360)	(0.963)
Type 6	0.00196	-0.00478	0.0245	-0.00880	0.00307
	(0.807)	(0.768)	(0.397)	(0.808)	(0.857)
Type 7	0.00910	-0.0356**	-0.0434	0.0547*	0.00406
	(0.284)	(0.028)	(0.215)	(0.060)	(0.838)
Type 8	-0.00432	-0.0218***	-0.0241**	0.0164	-0.0173
	(0.274)	(0.000)	(0.026)	(0.296)	(0.480)
Type 9	0.0128	0.0115**	0.0261*	-0.0368	0.00655
	(0.218)	(0.021)	(0.069)	(0.554)	(0.570)
Type 10	-0.0127	-0.0407	0.00564	-0.284**	0.0435
	(0.638)	(0.230)	(0.919)	(0.019)	(0.309)
Type 11	-0.00512	-0.0274	0.0287	-0.00517	-0.00665
	(0.680)	(0.218)	(0.545)	(0.931)	(0.725)
Type 12	-0.0229	0.00885	-0.0322	-0.168**	0.0827
	(0.150)	(0.329)	(0.485)	(0.024)	(0.307)
Type 13	0.0403***	0.0289***	0.0591***	0.138	0
	(0.000)	(0.000)	(0.001)	(0.156)	.
Type 14	-0.00557	0.00956	0.0445	0.0484	0.00987
	(0.706)	(0.410)	(0.234)	(0.736)	(0.611)
Type 15	-0.0207**	-0.00733	-0.0756***	-0.0551	0.155*
	(0.035)	(0.538)	(0.009)	(0.134)	(0.061)
Type 16	0.00835	0.00654	0.0248	-0.102	0.0304
	(0.568)	(0.608)	(0.183)	(0.358)	(0.100)
Type 17	0.000294	0.00952**	0.0124	-0.0607	0.0343***
	(0.981)	(0.035)	(0.344)	(0.167)	(0.009)
Type 18	0.00151	-0.000979	0.00389	0.0801***	-0.0179*
	(0.838)	(0.915)	(0.730)	(0.005)	(0.054)
Type 19	-0.00514	0.00841	0.00816	-0.00377	0.0182
	(0.554)	(0.101)	(0.522)	(0.909)	(0.191)
Type 20	0.00492	-0.000988	0.0109	-0.00547	0.00439
	(0.312)	(0.784)	(0.182)	(0.789)	(0.627)
Type 21	-0.0153*	0.00367	-0.0107	-0.00651	0.0151
	(0.078)	(0.468)	(0.446)	(0.805)	(0.134)
Type 22	-0.00796	0.0117**	0.00592	-0.0583*	0.0188
	(0.295)	(0.048)	(0.681)	(0.096)	(0.191)
Type 23	0.00215	-0.00282	-0.0184	-0.00563	0.00326
	(0.814)	(0.797)	(0.345)	(0.873)	(0.898)
Type 24	-0.00584	0.0198**	0.0508*	-0.0471	-0.106**
	(0.607)	(0.026)	(0.096)	(0.401)	(0.033)
Type 25	-0.0159	-0.00487	0.000804	-0.107	0.0138
	(0.413)	(0.598)	(0.971)	(0.300)	(0.356)
log_emp	-0.00103	0.00135	-0.00188	-0.00153	0.0202***
	(0.200)	(0.177)	(0.391)	(0.655)	(0.000)
N	10057	10057	10057	10041	3925
R^2	0.0346	0.0139	0.0207	0.121	0.0223

p-values in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Year Dummies and a constant are omitted.

Table 12 Predicting Risk Measures and Stock Returns by the First Ten Risk Factors

	(1) Option Volatility	(2) Debt Ratio	(3) Lowest Annual Stock Return	(4) Annual Stock Return	(5) Highest Annual Stock Return
Type 1	-0.00149	-0.00130	0.00291	0.0415*	0.0301
	(0.639)	(0.686)	(0.755)	(0.059)	(0.205)
Type 2	-0.00208	-0.0120**	-0.00472	-0.00989	-0.00354
	(0.735)	(0.048)	(0.816)	(0.836)	(0.940)
Type 3	0.00962**	0.00873**	-0.0291**	-0.00717	0.0147
	(0.027)	(0.042)	(0.017)	(0.808)	(0.661)
Type 4	-0.00330	0.00464*	-0.0102	-0.00719	0.000305
	(0.209)	(0.052)	(0.184)	(0.680)	(0.987)
Type 5	-0.00688***	-0.00146	-0.0115	-0.0742***	-0.0645***
	(0.009)	(0.574)	(0.149)	(0.000)	(0.001)
Type 6	0.00401	-0.00363	-0.0144	-0.0299	-0.0436*
	(0.282)	(0.175)	(0.172)	(0.201)	(0.070)
Type 7	0.0128**	-0.00411	-0.0214	-0.0489	-0.0186
	(0.011)	(0.263)	(0.105)	(0.111)	(0.573)
Type 8	0.00637**	-0.00758***	-0.0213***	-0.0284	-0.0315*
	(0.014)	(0.000)	(0.008)	(0.115)	(0.085)
Type 9	-0.00372	-0.00507**	-0.0225**	-0.0405*	-0.0629***
	(0.207)	(0.043)	(0.016)	(0.056)	(0.004)
Type 10	-0.00705**	-0.00284	-0.00552	-0.0217	-0.0235
	(0.037)	(0.453)	(0.612)	(0.388)	(0.377)
Type 11	0.00435	-0.00497	-0.00602	-0.0328	-0.0520**
	(0.204)	(0.110)	(0.571)	(0.182)	(0.035)
Type 12	-0.00576	0.00173	0.00298	0.0146	0.0101
	(0.154)	(0.622)	(0.801)	(0.590)	(0.721)
Type 13	0.00137	0.00978	-0.00243	-0.0199	0.00325
	(0.864)	(0.175)	(0.908)	(0.692)	(0.950)
Type 14	-0.00385	0.00149	0.00754	0.0104	0.0153
	(0.334)	(0.661)	(0.538)	(0.697)	(0.571)
Type 15	0.00184	-0.000469	0.0124	0.00253	-0.00371
	(0.517)	(0.853)	(0.130)	(0.892)	(0.854)
Type 16	0.00731	0.0193***	0.0173	0.0260	0.0384
	(0.200)	(0.005)	(0.312)	(0.502)	(0.387)
Type 17	-0.0114***	0.00161	0.00469	0.0105	0.00527
	(0.000)	(0.579)	(0.598)	(0.610)	(0.805)
Type 18	0.0101***	-0.00145	-0.0185*	-0.0153	-0.0371
	(0.003)	(0.617)	(0.063)	(0.500)	(0.111)
Type 19	-0.0117***	-0.00499*	-0.0204**	-0.0379*	-0.0482**
	(0.000)	(0.084)	(0.034)	(0.086)	(0.037)
Type 20	-0.00547**	0.00159	-0.00461	0.00960	0.0170
	(0.041)	(0.520)	(0.562)	(0.596)	(0.381)
Type 21	0.00201	-0.00431	-0.0126	0.00957	0.0308
	(0.578)	(0.158)	(0.279)	(0.716)	(0.265)
Type 22	-0.0117***	0.00107	0.0193**	0.0185	0.0109
	(0.000)	(0.706)	(0.034)	(0.375)	(0.618)
Type 23	0.00117	-0.000827	0.00712	0.0225	0.00960
	(0.745)	(0.804)	(0.516)	(0.363)	(0.721)
Type 24	-0.00939***	0.00587*	0.00134	-0.00203	-0.0164
	(0.002)	(0.096)	(0.893)	(0.930)	(0.511)
Type 25	0.00187	0.00416	0.00497	0.00136	-0.0162
	(0.657)	(0.294)	(0.715)	(0.963)	(0.613)
log_emp	0.00175**	-0.000908	-0.00356*	0.00437	-0.00258
	(0.013)	(0.176)	(0.082)	(0.352)	(0.616)
N	6559	10057	5958	5965	5929
R^2	0.191	0.0320	0.373	0.367	0.134

p-values in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Year Dummies and a constant are omitted.

Table 13 Predicting Financial Ratios by the First Ten Risk Factors

	(1) ROA	(2) Gross Margin	(3) Net Margin	(4) Tobin's Q	(5) R&D Ratio
Type 1	0.00785**	0.0165***	0.0292***	0.0168	-0.0683***
	(0.043)	(0.000)	(0.004)	(0.316)	(0.008)
Type 2	0.0177**	0.00359	0.0312**	0.00553	-0.0323
	(0.041)	(0.534)	(0.048)	(0.851)	(0.197)
Type 3	0.00357	-0.00987	0.0134	-0.00274	-0.0173
	(0.491)	(0.123)	(0.352)	(0.902)	(0.687)
Type 4	-0.00553*	-0.00678**	-0.0258***	0.00758	0.0116
	(0.065)	(0.030)	(0.001)	(0.557)	(0.482)
Type 5	-0.00165	-0.00291	-0.0192**	0.0225	-0.0203
	(0.610)	(0.377)	(0.019)	(0.110)	(0.268)
Type 6	-0.00831**	-0.00414	-0.0149*	0.0109	0.00895
	(0.016)	(0.391)	(0.099)	(0.489)	(0.642)
Type 7	-0.00143	-0.0221***	-0.0117	0.0178	0.0778*
	(0.753)	(0.000)	(0.395)	(0.350)	(0.063)
Type 8	0.00123	-0.0152***	-0.0260***	0.0233*	0.00488
	(0.676)	(0.000)	(0.000)	(0.055)	(0.631)
Type 9	-0.000595	0.00213	0.0000174	0.00569	0.0157
	(0.868)	(0.512)	(0.998)	(0.703)	(0.111)
Type 10	0.00906*	0.00564	0.0247**	-0.0555***	0.00158
	(0.056)	(0.211)	(0.029)	(0.009)	(0.935)
Type 11	0.00132	-0.000261	0.0169*	-0.00897	-0.00313
	(0.737)	(0.947)	(0.057)	(0.631)	(0.848)
Type 12	0.00342	0.000715	-0.00115	0.0134	0.0326**
	(0.465)	(0.846)	(0.915)	(0.492)	(0.037)
Type 13	0.00522	0.00992	0.00779	0.0634	0.0229
	(0.511)	(0.252)	(0.601)	(0.128)	(0.595)
Type 14	-0.00526	-0.00135	-0.00252	-0.0134	0.0337**
	(0.218)	(0.728)	(0.777)	(0.498)	(0.040)
Type 15	-0.00267	0.00244	0.00176	-0.0324**	-0.00284
	(0.399)	(0.466)	(0.822)	(0.023)	(0.877)
Type 16	-0.0149*	-0.0149*	-0.0171	-0.0169	0.00634
	(0.061)	(0.078)	(0.419)	(0.646)	(0.871)
Type 17	-0.00232	0.0126***	0.0188**	-0.0400**	-0.0155
	(0.517)	(0.000)	(0.017)	(0.012)	(0.372)
Type 18	-0.00231	0.00259	0.00267	0.0354**	-0.00724
	(0.508)	(0.442)	(0.708)	(0.016)	(0.524)
Type 19	-0.00998***	0.00102	-0.0166**	-0.00203	0.0472***
	(0.010)	(0.754)	(0.036)	(0.902)	(0.001)
Type 20	-0.00201	-0.00379	-0.0106	-0.00180	0.0199
	(0.513)	(0.261)	(0.180)	(0.894)	(0.275)
Type 21	-0.0121**	-0.000735	-0.0187**	0.0166	0.0193***
	(0.013)	(0.812)	(0.048)	(0.333)	(0.008)
Type 22	-0.00569	0.00566*	-0.00687	-0.0453***	0.0347**
	(0.133)	(0.079)	(0.396)	(0.004)	(0.020)
Type 23	0.00175	0.000185	-0.00788	-0.0239	0.0102
	(0.706)	(0.970)	(0.485)	(0.198)	(0.642)
Type 24	-0.00522	0.00897**	0.00269	-0.0122	-0.0357**
	(0.231)	(0.031)	(0.804)	(0.534)	(0.049)
Type 25	-0.00267	0.00488	-0.00200	-0.0328	0.0321**
	(0.625)	(0.202)	(0.847)	(0.200)	(0.036)
log_emp	-0.000541	0.00148	0.000218	-0.00504	0.0166***
	(0.526)	(0.135)	(0.922)	(0.152)	(0.001)
N	10057	10057	10057	10041	3925
R^2	0.0374	0.0201	0.0235	0.125	0.0254

p-values in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Year Dummies and a constant are omitted.

Table 14 Predicting Risk Measures and Stock Returns by All Risk Factors

	(1) Option Volatility	(2) Debt Ratio	(3) Lowest Annual Stock Return	(4) Annual Stock Return	(5) Highest Annual Stock Return
Type 1	0.00407	-0.00161	0.00104	0.0424**	0.0566***
	(0.137)	(0.534)	(0.899)	(0.025)	(0.005)
Type 2	0.00570	-0.00904	-0.00433	-0.0305	-0.0404
	(0.247)	(0.106)	(0.791)	(0.436)	(0.319)
Type 3	0.00796**	0.00203	-0.0288***	0.0146	0.0249
	(0.012)	(0.508)	(0.002)	(0.503)	(0.298)
Type 4	0.0000639	0.00517**	-0.0177**	-0.0120	0.00574
	(0.982)	(0.048)	(0.033)	(0.526)	(0.777)
Type 5	0.00272	0.00657	0.00512	0.00822	0.00612
	(0.628)	(0.116)	(0.706)	(0.797)	(0.859)
Type 6	0.00711***	-0.00188	0.00717	0.0146	0.0111
	(0.009)	(0.429)	(0.365)	(0.421)	(0.558)
Type 7	0.000724	0.00229	-0.00593	0.000880	0.00516
	(0.800)	(0.387)	(0.479)	(0.964)	(0.803)
Type 8	0.00942***	-0.00483**	-0.0213***	-0.00441	-0.00115
	(0.000)	(0.034)	(0.006)	(0.802)	(0.950)
Type 9	0.00175	-0.00696***	-0.0175**	-0.0243	-0.0428**
	(0.517)	(0.008)	(0.031)	(0.195)	(0.030)
Type 10	-0.0170***	0.00548*	0.00911	0.0234	0.0270
	(0.000)	(0.065)	(0.334)	(0.280)	(0.242)
Type 11	-0.000829	-0.0000948	0.00836	0.00628	-0.00804
	(0.763)	(0.971)	(0.305)	(0.742)	(0.690)
Type 12	-0.00364	-0.000157	0.00306	0.00817	-0.00405
	(0.164)	(0.953)	(0.721)	(0.679)	(0.844)
Type 13	0.00832*	0.00773	0.0121	0.0180	0.0317
	(0.067)	(0.101)	(0.398)	(0.593)	(0.364)
Type 14	-0.00448	0.000762	0.0206**	0.0371*	0.0462**
	(0.103)	(0.770)	(0.016)	(0.057)	(0.024)
Type 15	0.00121	-0.000683	0.00327	-0.0132	-0.0127
	(0.692)	(0.790)	(0.709)	(0.497)	(0.529)
Type 16	0.00223	0.00957**	0.00733	0.00864	0.0454
	(0.534)	(0.046)	(0.550)	(0.762)	(0.153)
Type 17	-0.00535*	-0.000303	0.0128	0.0409**	0.0163
	(0.062)	(0.908)	(0.126)	(0.032)	(0.423)
Type 18	0.00335	0.000313	-0.0150*	-0.000184	-0.00875
	(0.239)	(0.905)	(0.083)	(0.993)	(0.674)
Type 19	-0.00718**	-0.00165	-0.0112	-0.0122	-0.0190
	(0.013)	(0.552)	(0.208)	(0.549)	(0.376)
Type 20	0.00361	0.00555*	-0.00904	0.00608	0.0144
	(0.268)	(0.071)	(0.371)	(0.790)	(0.559)
Type 21	-0.000576	-0.00566**	-0.0138	-0.00162	0.0207
	(0.863)	(0.046)	(0.187)	(0.947)	(0.414)
Type 22	-0.00931***	0.00146	0.0181**	0.0309	0.0171
	(0.001)	(0.600)	(0.032)	(0.114)	(0.408)
Type 23	-0.00220	-0.000590	0.00739	0.0120	0.0176
	(0.466)	(0.814)	(0.393)	(0.533)	(0.380)
Type 24	-0.00160	0.00300	-0.000992	0.00626	0.00146
	(0.618)	(0.364)	(0.920)	(0.784)	(0.953)
Type 25	0.000615	0.00132	0.00103	0.00780	0.00643
	(0.857)	(0.701)	(0.926)	(0.750)	(0.804)
log_emp	0.0000480	-0.000795	-0.00407*	0.00319	-0.00285
	(0.949)	(0.236)	(0.056)	(0.511)	(0.590)
N	6559	10057	5958	5965	5929
R^2	0.194	0.0312	0.375	0.366	0.133

p-values in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Year Dummies and a constant are omitted.

Table 15 Predicting Financial Ratios by All Risk Factors

	(1) ROA	(2) Gross Margin	(3) Net Margin	(4) Tobin's Q	(5) R&D Ratio
Type 1	0.00783**	0.0110***	0.0185**	0.0485***	-0.0472**
	(0.016)	(0.001)	(0.023)	(0.000)	(0.013)
Type 2	0.0102	0.00194	0.0223	0.0337	-0.0651*
	(0.162)	(0.721)	(0.165)	(0.201)	(0.054)
Type 3	0.00438	-0.00814*	0.00233	-0.0138	0.0217
	(0.254)	(0.055)	(0.819)	(0.412)	(0.383)
Type 4	-0.00478	-0.00385	-0.0183**	-0.0131	-0.00412
	(0.127)	(0.270)	(0.020)	(0.360)	(0.840)
Type 5	0.00844	-0.00787**	-0.00440	0.0634***	-0.0438
	(0.125)	(0.033)	(0.708)	(0.009)	(0.108)
Type 6	-0.00447	-0.00815**	-0.0107	0.0108	0.00699
	(0.149)	(0.016)	(0.180)	(0.415)	(0.708)
Type 7	0.00120	-0.00480	-0.000437	0.0202	-0.0259
	(0.717)	(0.167)	(0.958)	(0.167)	(0.219)
Type 8	-0.000731	-0.0134***	-0.0247***	0.0228*	0.0129
	(0.811)	(0.000)	(0.000)	(0.076)	(0.309)
Type 9	0.0000649	-0.000906	-0.00408	-0.00522	0.0553***
	(0.984)	(0.792)	(0.618)	(0.720)	(0.003)
Type 10	0.00152	0.00637*	0.0190**	-0.0666***	-0.00568
	(0.683)	(0.074)	(0.033)	(0.000)	(0.653)
Type 11	0.0000187	0.00500	0.0244***	-0.00748	-0.0305**
	(0.995)	(0.132)	(0.002)	(0.599)	(0.043)
Type 12	0.00954***	0.00392	0.0150*	0.00783	-0.00249
	(0.006)	(0.249)	(0.070)	(0.598)	(0.883)
Type 13	0.00140	-0.000433	0.00514	-0.0182	0.0550*
	(0.795)	(0.935)	(0.667)	(0.471)	(0.078)
Type 14	-0.00350	-0.000145	-0.00447	0.000466	0.00828
	(0.302)	(0.966)	(0.566)	(0.975)	(0.627)
Type 15	0.0000384	0.00414	0.0104	-0.0309**	-0.0387*
	(0.990)	(0.254)	(0.188)	(0.026)	(0.062)
Type 16	-0.00337	-0.00139	0.00454	-0.0102	0.00942
	(0.554)	(0.798)	(0.742)	(0.691)	(0.721)
Type 17	0.00478	0.0121***	0.0272***	-0.0154	0.00725
	(0.142)	(0.000)	(0.000)	(0.285)	(0.717)
Type 18	-0.00292	0.00641*	0.0127*	0.000451	-0.00120
	(0.355)	(0.058)	(0.082)	(0.975)	(0.926)
Type 19	-0.00812**	0.00377	-0.0151*	0.0120	0.0477***
	(0.025)	(0.248)	(0.057)	(0.433)	(0.000)
Type 20	-0.00424	-0.00864**	-0.0253**	0.00484	0.0406
	(0.245)	(0.045)	(0.013)	(0.776)	(0.160)
Type 21	-0.0119***	-0.000429	-0.0201**	0.00520	0.0193**
	(0.006)	(0.888)	(0.025)	(0.750)	(0.021)
Type 22	-0.000642	0.00732**	-0.00413	-0.0141	0.0699***
	(0.847)	(0.038)	(0.627)	(0.328)	(0.001)
Type 23	-0.00324	0.00399	0.00747	-0.0282**	0.00101
	(0.305)	(0.223)	(0.327)	(0.035)	(0.946)
Type 24	-0.00286	0.00701*	-0.00107	0.00808	-0.0138
	(0.499)	(0.073)	(0.913)	(0.664)	(0.420)
Type 25	-0.00359	0.000255	-0.00380	-0.0145	-0.00121
	(0.429)	(0.950)	(0.717)	(0.464)	(0.948)
log_emp	-0.00102	0.00203**	0.000758	-0.00693*	0.0166***
	(0.234)	(0.050)	(0.740)	(0.059)	(0.004)
N	10057	10057	10057	10041	3925
R^2	0.0369	0.0210	0.0255	0.127	0.0327

p-values in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Year Dummies and a constant are omitted.

Table 16 The Number of Stars from Table 10 to Table 15

Type	Option Volatility	Debt Ratio	Lowest Return	1-Year Return	Highest Return	ROA	Gross Margin	Net Margin	Tobin's Q	R&D Ratio	Total	
1	2	0	0	3	3	5	9	8	3	8	41	1st
2	2	2	0	0	0	2	0	2	0	1	9	
3	7	2	5	0	0	0	1	0	0	0	15	
4	0	3	2	0	2	1	3	5	0	0	16	
5	5	1	0	3	4	2	2	2	3	0	22	
6	3	0	0	1	3	2	2	1	0	0	12	
7	3	2	0	0	0	0	5	0	1	1	12	
8	5	5	8	2	1	0	9	8	2	0	40	2nd
9	0	5	4	1	5	0	2	1	0	3	21	
10	5	1	0	0	0	1	1	4	8	0	20	
11	0	2	0	0	2	0	0	4	0	2	10	
12	1	0	0	0	0	3	0	1	2	2	9	
13	1	3	0	3	2	3	3	3	0	1	19	
14	1	0	2	1	2	0	0	0	0	2	8	
15	0	0	0	0	0	2	0	3	4	2	11	
16	0	5	0	0	0	1	1	0	0	1	8	
17	7	2	0	2	0	0	8	5	2	3	29	3rd
18	6	0	2	0	0	0	1	1	5	1	16	
19	8	1	2	1	2	5	0	3	0	6	28	4th
20	5	1	0	0	0	0	2	2	0	0	10	
21	2	2	2	0	0	6	0	4	0	5	21	
22	9	0	5	0	0	0	5	0	4	5	28	4th
23	0	1	0	0	0	0	0	0	2	0	3	Lowest
24	3	2	0	0	0	0	5	1	0	4	15	
25	0	3	0	3	0	3	0	3	0	5	17	
	75	43	32	20	26	36	59	61	36	52	440	

References

1. Antweiler, W., and Frank, M.Z. 2004. "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards," *Journal of Finance* (59:3), Jun, pp 1259-1294.
2. Baesens, B., Setiono, R., Mues, C., and Vanthienen, J. 2003. "Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation," *Management Science* (49:3), Mar, pp 312-329.
3. Bai, X. 2010. "Predicting Consumer Sentiments from Online Text," *Decision Support Systems* (Forthcoming).
4. Bai, X., Nunez, M., and Kalagnanam, J. 2010. "Managing Data Quality Risk in Accounting Information Systems," *Information Systems Research* (Forthcoming).
5. Balakrishnan, R., Qiu, X.Y., and Srinivasan, P. 2010. "On the Predictive Ability of Narrative Disclosures in Annual Reports," *European Journal of Operational Research* (202:3), May 1, pp 789-801.
6. Barber, B., and Lyon, J. 1996. "Detecting Abnormal Operating Performance: The Empirical Power and Specification of Test Statistics," *Journal of financial Economics* (41:3), pp 359-399.
7. Barua, A., Konana, P., Whinston, A.B., and Yin, F. 2004. "An Empirical Investigation of Net-Enabled Business Value," *Mis Quarterly* (28:4), Dec, pp 585-620.
8. Bharadwaj, A.S. 2000. "A Resource-Based Perspective on Information Technology Capability and Firm Performance: An Empirical Investigation," *Mis Quarterly* (24:1), Mar, pp 169-196.
9. Bharadwaj, A.S., Bharadwaj, S.G., and Konsynski, B.R. 1999. "Information Technology Effects on Firm Performance as Measured by Tobin's Q," *Management Science* (45:7), Jul, pp 1008-1024.
10. Cecchini, M., Aytug, H., Koehler, G.J., and Pathak, P. 2010. "Detecting Management Fraud in Public Companies," *Management Science* (56:7), Jul, pp 1146-1160.
11. Das, S., and Chen, M. 2007. "Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web," *Management Science* (53:9), pp 1375-1388.
12. Feldman, R., Govindaraj, S., Livnat, J., and Segal, B. 2009. "Management's Tone Change, Post Earnings Announcement Drift and Accruals," *Review of Accounting Studies* (Forthcoming).
13. Gu, B., Konana, P., Rajagopalan, B., and Chen, H. 2007. "Competition among Virtual Communities and User Valuation: The Case of Investing-Related Communities," *Information Systems Research* (18:1), Mar, pp 68-85.
14. Han, J., and Kamber, M. 2006. *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
15. Hanley, K., and Hoberg, G. 2010. "The Information Content of IPO Prospectuses," *Review of Financial Studies* (23:7), pp 2821-2864.
16. Huang, Z., Chen, H.C., Hsu, C.J., Chen, W.H., and Wu, S.S. 2004a. "Credit Rating Analysis with Support Vector Machines and Neural Networks: A Market Comparative Study," *Decision Support Systems* (37:4), Sep, pp 543-558.
17. Huang, Z., Chung, W.Y., and Chen, H.C. 2004b. "A Graph Model for E-Commerce Recommender Systems," *Journal of the American Society for Information Science and Technology* (55:3), Feb, pp 259-274.
18. Huang, Z., Zeng, D.D., and Chen, H.C. 2007. "Analyzing Consumer-Product Graphs: Empirical Findings and Applications in Recommender Systems," *Management Science* (53:7), Jul, pp 1146-1164.
19. Hull, J. 2008. *Options, Futures, and Other Derivatives*. Pearson Prentice Hall.
20. Kothari, S.P., Li, X., and Short, J.E. 2009. "The Effect of Disclosures by Management, Analysts, and Business Press on Cost of Capital, Return Volatility, and Analyst Forecasts: A Study Using Content Analysis," *Accounting Review* (84:5), Sep, pp 1639-1670.

21. Li, F. 2008. "Annual Report Readability, Current Earnings, and Earnings Persistence," *Journal of Accounting and Economics* (45:2-3), pp 221-247.
22. Li, F. 2010. "The Information Content of Forward-Looking Statements in Corporate Filings—a Naive Bayesian Machine Learning Approach," *Journal of Accounting Research* (Forthcoming).
23. Loughran, T., and McDonald, B. 2009. "When Is a Liability Not a Liability," *Journal of Finance* (Forthcoming).
24. Loughran, T., and McDonald, B. 2010. "Measuring Readability in Financial Text," *SSRN* (working papers).
25. Mangen, C., and Durnev, A. 2010. "The Real Effects of Disclosure Tone: Evidence from Restatements," *SSRN* (working papers).
26. Padmanabhan, B., and Tuzhilin, A. 2003. "On the Use of Optimization for Data Mining: Theoretical Interactions and eCRM Opportunities," *Management Science* (49:10), Oct, pp 1327-1343.
27. Padmanabhan, B., Zheng, Z.Q., and Kimbrough, S.O. 2006. "An Empirical Analysis of the Value of Complete Information for eCRM Models," *Mis Quarterly* (30:2), Jun, pp 247-267.
28. Provost, F., and Fawcett, T. 2001. "Robust Classification for Imprecise Environments," *Machine Learning* (42:3), pp 203-231.
29. Sarkar, S., and Sriram, R.S. 2001. "Bayesian Models for Early Warning of Bank Failures," *Management Science* (47:11), Nov, pp 1457-1475.
30. Schumaker, R.P., and Chen, H.C. 2009. "Textual Analysis of Stock Market Prediction Using Breaking Financial News: The Azfintext System," *Acm Transactions on Information Systems* (27:2), pp -.
31. Tetlock, P., Saar-Tsechansky, M., and Macskassy, S. 2008. "More Than Words: Quantifying Language to Measure Firms' Fundamentals," *The Journal of Finance* (63:3), pp 1437-1467.
32. Tetlock, P.C. 2007. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market," *Journal of Finance* (62:3), Jun, pp 1139-1168.
33. Watson, C. 1990. "Multivariate Distributional Properties, Outliers, and Transformation of Financial Ratios," *Accounting Review* (65:3), pp 682-695.
34. Wei, C., Chen, Y., Yang, C., and Yang, C. 2010. "Understanding What Concerns Consumers: A Semantic Approach to Product Feature Extraction from Consumer Reviews," *Information Systems and E-Business Management* (8:2), pp 149-167.
35. Wooldridge, J. 2002. *Econometric Analysis of Cross Section and Panel Data*. The MIT press.
36. Wooldridge, J. 2009. *Introductory Econometrics: A Modern Approach*. South Western Cengage Learning.