Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures

Yang Bao, Anindya Datta

To cite this article:
Yang Bao, Anindya Datta (2014) Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures. Management Science 60(6):1371-1391. http://dx.doi.org/10.1287/mnsc.2014.1930

MANAGEMENT SCIENCE
Vol. 60, No. 6, June 2014, pp. 1371–1391
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2014.1930

Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures

Yang Bao, Anindya Datta
Department of Information Systems, National University of Singapore, Singapore 117417
{baoyang9527@gmail.com, datta@comp.nus.edu.sg}

Managers and researchers alike have long recognized the importance of corporate textual risk disclosures. Yet it is a nontrivial task to discover and quantify variables of interest from unstructured text. In this paper, we develop a variation of the latent Dirichlet allocation topic model and its learning algorithm for simultaneously discovering and quantifying risk types from textual risk disclosures. We conduct comprehensive evaluations in terms of both conventional statistical fit and substantive fit with respect to the quality of discovered information. Experimental results show that our proposed method outperforms all competing methods, and could find more meaningful topics (risk types). By taking advantage of our proposed method for measuring risk types from textual data, we study how risk disclosures in 10-K forms affect the risk perceptions of investors. Different from prior studies, our results provide support for all three competing arguments regarding whether and how risk disclosures affect the risk perceptions of investors, depending on the specific risk types disclosed. We find that around two-thirds of risk types lack informativeness and have no significant influence. Moreover, we find that the informative risk types do not necessarily increase the risk perceptions of investors—the disclosure of three types of systematic and liquidity risks will increase the risk perceptions of investors, whereas the other five types of unsystematic risks will decrease them.

Data, as supplemental material, are available at http://dx.doi.org/10.1287/mnsc.2014.1930.

Keywords: topic modeling; latent Dirichlet allocation; text analysis; econometric analysis; risk disclosures
History: Received September 16, 2012; accepted January 12, 2014, by Alok Gupta, special issue on business analytics. Published online in *Articles in Advance* April 11, 2014.

1. Introduction

The annual report issued by a corporation is an important source of information for its stakeholders, such as investors, to obtain a detailed picture of the company's business, the risks it faces, and its operating and financial results. The filing of annual reports is typically mandated by the relevant regulatory agency in the country of the corporation's domicile. Most U.S. public companies, for example, are required by the U.S. Securities and Exchange Commission (SEC) to issue an annual report in a well-defined format (specified in the SEC 10-K form), each year. In addition to the quantitative financial data detailed in these reports, one of the most analyzed elements in the 10-K form is the risk disclosures about the corporation, since stakeholders are particularly sensitive to risks. These risk disclosures are considered so important, that starting in 2005, the SEC required that all firms include a separate section (section 1A) in their 10-K form to discuss "the most significant factors that make the offering speculative or risky" (SEC 2005, regulation S-K, item 503(c)). This section has turned out to be one of the most examined and debated segments of corporate annual reports (Campbell et al. 2014).

Conceptually, there are many potential variables of interest from these risk disclosures. Some of these variables, as surveyed by Li (2010b), include the *amount*, *tone*, and *transparency* (or readability) of disclosures. In this paper, we are particularly interested in another important variable, the *risk type*, which has been paid less attention than the variables identified previously (Mirakur 2011, Campbell et al. 2014, Huang and Li 2011). At a high level, risk types refer to general factors that present elements of risk to a corporation, such as litigation, or natural disasters. We should also note that the risk disclosure section in an annual report appears as a free-form textual segment, i.e., completely unstructured text.

Discovering and quantifying variables of interest (such as risk types discussed in this paper) from large amounts of unstructured text is a nontrivial task for social science researchers. They have struggled when confronted with this problem, since it is difficult, indeed infeasible, to manually perform exhaustive text perusal, even in a moderately sized corpus. For example, Mirakur (2011) has manually categorized 29 risk types for 122 randomly selected firms. This sample is far less than 1% of the total number of published 10-K forms.

In this scenario, it is tempting to apply *automated text analysis* to this important problem. Indeed, researchers have gone down this path: Campbell et al. (2014) use a predefined dictionary to quantify five risk types in 10-K forms: *idiosyncratic*, *systematic*, *financial*, *tax*, and *litigation* risks. Huang and Li (2011) propose a supervised learning method to automatically categorize risk factors reported in section 1A of 10-K forms into 25 risk types. In §2 we will provide a comprehensive review of this work. Suffice it to mention now that all of this work falls into two categories of automated text analysis: *dictionary* based and *supervised learning* based.

Dictionary and supervised learning methods assume a predefined set of categories. This assumption poses no challenge if researchers have a set of categories for texts in mind. For example, if researchers aim to identify positive and negative tones of textual statements (a common theme of work), the categories are quite explicit (i.e., *positive* and *negative*). In most cases, however, the categories might be hard to derive beforehand. Take our case for example. The risk factors affecting firms are (a) unpredictable and (b) differ from firm to firm. Clearly, a priori knowledge of what a corporation might perceive as risk is impossible to achieve. Without this knowledge, it would be impossible to apply dictionary or supervised learning methods to identify what types of risks are disclosed in section 1A of 10-K forms. Unfortunately, all prior work is based on the notion of *predefined risk types*. The drawback of this assumption is further indicated by the salient difference between predefined risk types defined in Mirakur (2011), Campbell et al. (2014), and Huang and Li (2011). What is clearly needed is not only the ability to *quantify* risk types, but also to *discover* these risk types.

To bridge this gap, in this paper, we report the first general work on simultaneously discovering and quantifying risk types from textual risk disclosures. Specifically, we propose an unsupervised learning method that aims to estimate rather than predefine a set of categories (risk types) and simultaneously assign sentences (risk factors) to those categories. Quantification can then be easily achieved by aggregating counts for each document across risk types. To get a feel for our solution, consider Apple Inc.'s 10-K form in 2006.[1] From section 1A of this form, the summary headings of three sample risk factors are listed in Table 1. We assume that each risk factor only discusses one risk type. Our proposed method would take all the disclosed risk factors as input and yield a set of risk types and then map each risk factor to a risk type. For instance, for the factors disclosed in Table 1, our method would yield the following risk types: lawsuits (RT1), catastrophes (RT2), and human resources (RT3). Furthermore, it would map the first risk factor in Table 1 to RT1, the second factor to RT2, and the last factor to RT3. Finally, this disclosure document (if it only contains these three risk factors) could be quantified as a vector $[1, 1, 1]$, where each dimension corresponds to a risk type.

Table 1 Three Sample Risk Factors Disclosed in Section 1A of Apple Inc.'s 10-K Form in 2006

The matters relating to the investigation by the special committee of the board of directors and the restatement of the company's consolidated financial statements may result in additional litigation and governmental enforcement actions.
War, terrorism, public health issues, and other circumstances could disrupt supply, delivery, or demand of products, which could negatively affect the company's operations and performance.
The company's success depends largely on its ability to attract and retain key personnel.

Now we provide a short overview of our approach. We have designed a class of unsupervised clustering methods, using the technique of *topic modeling*. Although unsupervised methods allow us to strive for our goal of risk discovery, they come with their own intrinsic issues. Particularly, unsupervised clustering methods are generally application independent and may not be able to find the clusters (categories) that are "meaningful" for the task in a specific context. For resolving this problem, we extend the standard latent Dirichlet allocation (LDA) topic model, proposed by Blei et al. (2003), by taking into account the sentence structure—we assume that the words in each sentence (risk factor) are concerned with only one risk type. This additional contextual information will relax the bag-of-words assumption of the standard LDA model, and yield, as demonstrated conclusively by our experiments, more meaningful risk types.

It is worth noting that a recent debate, mainly in political science, casts unsupervised and supervised methods as competitor methods. We disagree. Rather, we subscribe to the view of Grimmer and Stewart (2013, p. 281) that "Far from competitors, supervised and unsupervised methods are most productively viewed as complimentary methods, particularly for new projects or recently collected data." Specifically, the categories of interest in a new corpus are usually unclear. In this case, unsupervised methods provide insights into category taxonomy that would be difficult to obtain. Once the unsupervised method is fit, supervised learning methods could be used to validate or generalize the findings.

By taking advantage of our proposed method for measuring risk types from textual data, we conduct an empirical study to investigate the effects of risk disclosures on the postdisclosure risk perceptions of investors. Our proposed method facilitates the investigation of the effects of risk disclosures at the individual risk type level, a previously tedious task.

[1] http://www.sec.gov/Archives/edgar/data/320193/0001104659-06-084288.txt (accessed March 22, 2014).

The main contributions of this paper are summarized as follows.

- We propose a novel unsupervised topic model, called sent-LDA (coupled with the variational expectation maximization (EM) learning algorithm), for simultaneously discovering and quantifying risk types from unstructured textual corporate risk disclosures. As far as we know, this is the first work that introduces unsupervised learning methods into the field of financial accounting.
- We conduct a comprehensive evaluation of our proposed model in terms of both conventional statistical fit and substantive fit with respect to the quality of discovered information. Experimental results show that our proposed method outperforms all competing methods and discovers more meaningful topics (risk types). We further visualize our learned model in a publicly available system[2]—the reader is welcome to try this out. The system facilitates the navigation of large amounts of textual risk disclosures by our target user base, including financial analysts, business managers, or academic researchers.
- Our proposed sent-LDA topic model (with the variational EM learning algorithm) contributes to the automated text analysis literature as well. It provides a flexible framework for incorporating contextual information (features of sentence structure in our case) to unearth more meaningful topics. Although our method is utilized for discovering and quantifying risk types from textual disclosures in this paper, it is general enough to be applicable to problems in other application domains as long as our key assumption that each sentence discusses only one topic is not violated. One significant possible application is the textual analysis of user generated online reviews where each sentence in the review only discusses one aspect of the target item.
- Our empirical study on the effect of disclosures of different risk types on the postdisclosure risk perception of investors contributes to the risk disclosure literature. Different from prior studies, our results provide support for all three competing arguments regarding whether and how risk disclosures (in section 1A of the 10-K form) affect the risk perceptions of investors, depending on the specific risk types disclosed. We find that around two-thirds of risk types lack informativeness and have no significant influence. Moreover, we find that the informative risk types do not necessarily increase the risk perceptions of investors—the disclosure of three types of systematic and liquidity risks will increase the risk perceptions of investors, whereas the other five types of unsystematic risks will decrease them.

[2] http://www.comp.nus.edu.sg/~baoyang/10kslda/browse/topic-list.html (accessed March 22, 2014).

2. Literature Review

This work is related to two primary areas: (1) automated text analysis and (2) effects of corporate risk disclosures.

2.1. Automated Text Analysis

Our work belongs to the well-known research area of automated text analysis, which aims to quantify textual information. As surveyed by O'Connor et al. (2011), there is an increasing interest in the use of automated text analysis in the services of social science questions. They argue that automated text analysis, which draws on techniques developed in natural language processing, information retrieval, text mining, and machine learning, should be properly understood as a class of quantitative social science methodologies. Although still in its growing stage, automated text analysis has been applied in many fields of social science, including political science (Grimmer 2010), economics (Aral et al. 2011), psychology (Tausczik and Pennebaker 2010), and others.

The most common use of automated text analysis in social science is to assign texts to categories. After categorizing, texts can be easily quantified using aggregated counts of categories. For example, to investigate the effects of risk disclosures, researchers may be interested in what risk types (e.g., potential lawsuits, catastrophes, etc.) are disclosed in corporate annual reports. In this example, the goal is to categorize each unit (e.g., word or sentence) of each annual report into one or more risk types and to aggregate counts across categories for quantifying each document.

Manual categorization is very resource (personnel, time) consuming. Even if the coding rules are developed and coders are trained, coders are still required to read each individual document. Automated text analysis could mitigate the cost of manual categorization, by limiting the amount of human effort. There are broadly three sorts of methods for automated text analysis, including (1) *dictionary*, (2) *supervised learning*, and (3) *unsupervised learning*.

2.1.1. Dictionary Methods. Dictionary methods are the most simple and intuitive automated text analysis methods. They use keywords or phrases to classify documents into categories or measure the extent to which documents belong to a particular category.

Because of its simplicity, dictionary methods are widely used for measuring texts in social science. Taking risk disclosure text analysis for example, Kothari et al. (2009) use dictionary methods to calculate the number of positives and negatives in each disclosure text, for each of the six business categories defined, by business quarter, by source. Feldman et al. (2010) use a classification scheme of words into positive and negative categories to measure the tone change in the MD&A section relative to prior periodic SEC filings. Campbell et al. (2014) create a keyword list by risk

category based on the dictionaries used in prior works and then use this list for classifying risk disclosures in section 1A of 10-K forms into five categories, including systematic, idiosyncratic, financial, tax, and legal risks. Rogers et al. (2011) use both general-purpose and context-specific text dictionaries to quantify optimistic and pessimistic tones in firms' earnings announcements. Kravet and Muslu (2011) extract risk disclosures from 10-K form by searching for sentences that involve the predefined risk-related keywords. They further categorize risk disclosure sentences to a negative tone if they contain one or more predefined keywords or their variations.

Dictionary methods require researchers to identify words that separate categorizations beforehand. In other words, researchers have to decide how categories should be assigned to documents using the defined dictionary. This may lead to inefficiencies when the dictionaries are applied outside the domain they were originally developed (Li 2010a).

2.1.2. Supervised Learning Methods. Supervised learning methods provide an alternative method for assigning documents to predefined categories. The idea is that (1) human coders first categorize a set of documents by hand, (2) then they train a supervised model that automatically learns how to assign categories to documents using coded data (training set). Supervised learning methods have two major advantages over dictionary methods (Grimmer and Stewart 2013). First, it is necessarily domain specific and therefore avoids the problems of applying dictionaries outside their intended area of use. Specifically, researchers have to develop coding rules for the variables (categories) of interest, forcing them to be clear about the definition and measurement of those variables. Second, they are easy to be validated using clear performance statistics.

Owing to its advantages, supervised learning methods have recently been applied in a number of social science contexts. Taking risk disclosure text analysis for example, Li (2010a) uses a naive Bayesian classifier to classify the tone and content of forward-looking statements in corporate 10-K and 10-Q filings. Huang and Li (2011) develop a multilabel text classification algorithm to classify risk factors in section 1A of 10-K form into 25 risk types. Humpherys et al. (2011) propose to use linguistic features to distinguish fraudulent from nonfraudulent 10-K reports using off-the-shelves classifiers. Cecchini et al. (2010) develop a method for automatically creating ontology for texts in MD&A section of 10-K forms, which could then be used for classifying financial events of firms.

2.1.3. Unsupervised Learning Methods. Dictionary and supervised learning methods assume a predefined set of categories. In contrast, unsupervised learning methods are a class of methods that learn underlying features of text without explicitly imposing categories of interests. They are usually called unsupervised clustering methods, where "clustering" means unsupervised "categorization." Unsupervised clustering methods use modeling assumptions and properties of the texts to estimate a set of categories and simultaneously assign documents (or other units of analysis such as sentences) to those categories. They are valuable since they could identify organizations of texts that are theoretically useful but perhaps understudied or previously unknown (Grimmer and Stewart 2013).

The problem of unsupervised clustering methods, as indicated by Grimmer and King (2011), is that they require a single, precisely defined objective function that works across applications. This is infeasible given that human beings are typically optimizing a goal of "useful" conceptualizations in an application-specific context. Grimmer and Stewart (2013) point out that there are two strategies to tackle this problem. The first strategy is to allow researchers to efficiently search over the potential categorization schemes for identifying interesting or useful organizations of the texts. For example, Grimmer and King (2011) develop a computer-assisted method for the discovery of insightful conceptualizations in the form of clusterings of input objects. The second strategy is to incorporate context-specific structure into the analysis through a model. The inclusion of this additional information often leads to more interesting clustering, but needs the variation of models. For example, Grimmer (2010) extends the LDA topic model by including the author information for better measuring the priorities political actors articulate in texts. Different from LDA models, which assumes the document-specific topic mixture, the extended model assumes that all documents of the same author share the same topic mixture.

The use of unsupervised clustering methods to analyze texts in social science is still at its infancy. Most attempts, such as the works of Grimmer and King (2011) and Grimmer (2010) mentioned previously, are in the field of political science. Additionally, Aral et al. (2011) use unsupervised topic models to characterize the topical content of stock recommendation articles. However, they directly use the standard LDA model and do not make any modification for incorporating context-specific information in order to solve the aforementioned problem of unsupervised clustering methods.

As far as we know, no previous works attempt to use unsupervised clustering methods for analyzing corporate risk disclosures. We report the first work to simultaneously discover the topics (risk types) in the data, assign sentences (risk factors) to their likely topics, and quantify the attention each disclosure document dedicates to the estimated topics. Our proposed sent-LDA topic model resolves the aforementioned problem of unsupervised clustering by taking into

account the contextual information about sentence structure—words in each sentence (risk factor) are probably discussing only one risk type.

2.2. Effects of Corporate Risk Disclosures

Our work is also related to studies on the effects of corporate risk disclosures. As surveyed by Li (2010b), researchers have long recognized the importance of understanding corporate risk disclosures and conducted empirical studies to investigate various research questions. Here, we particularly focus on the effects of corporate risk disclosures on investors' postdisclosure risk perceptions.

There are competing arguments about whether and how risk disclosures affect users' risk perceptions. The first argument is that risk disclosures are by and large boilerplate (*null argument*). There is a long-standing criticism that risk disclosures in financial reports are unlikely to be informative (Schrand and Elliott 1998). The critics argue that the managers are likely to disclose all possible risks and uncertainties without considering their impacts on firms, and thus the disclosed risks are vague and boilerplate in nature.

The second argument is that risk disclosures reveal previously unknown risk factors and contingencies, thereby increasing users' risk perceptions (*divergence argument*). For example, Campbell et al. (2014) find that the lengths of section 1A in 10-K forms (in which companies state their risk factors) are associated with low bid-ask spreads (a proxy for information asymmetry) and high beta and stock return volatility (a proxy for investors' assessments of fundamental risk) in the following year. Kravet and Muslu (2011) find that annual increases in risk disclosures are associated with increased stock return volatility and trading volume around and after the filings, suggesting that textual risk disclosures increase investors' risk perceptions.

The third argument is that risk disclosures resolve a firm's known risk factors and contingencies, thereby reducing users' risk perception (*convergence argument*). For example, Rajgopal (1999) finds that oil and gas firms' disclosures about market exposures are associated with stock return sensitivities to oil and gas prices. Linsmeier et al. (2002) find that after firms disclose mandated information about their exposures to interest rates, foreign currency exchange rates, and energy prices, trading volume sensitivity to changes in these underlying market rates and prices declines, even after controlling for other factors associated with trading volume.

Kothari et al. (2009) argue that the mixed findings of previous empirical studies are because disclosure tone affects the relationship between disclosure and measures of a firm's capital market environment (e.g., cost of capital, return volatility, etc.). Specifically, favorable disclosures result in lower return volatility whereas unfavorable disclosures lead to higher return volatility.

In contrast to previous literature, we further hypothesize the mixed findings are due to the semantic content of disclosures. Specifically, the *direction* of the relation between disclosures and users' risk perceptions depends on the disclosed risk types, which are discovered and quantified using our proposed method. We limit our analysis to newly created risk disclosures, i.e., section 1A in 10-K forms, since we know that the tone of this section is negative/pessimistic (Campbell et al. 2014). This allows us to test our hypotheses with the control of disclosure tone.

3. Proposed Methods

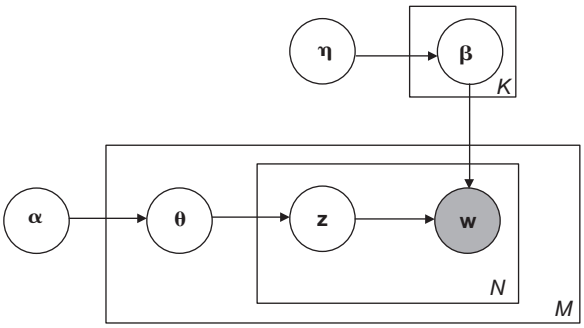
In this section, we propose to use unsupervised topic models for discovering and quantifying risk types from textual risk disclosures. We first introduce the LDA model, and then describe the intuitions behind our extension of LDA to discover more meaningful topics for representing risk types. After that, we elaborate our proposed sent-LDA model and its learning algorithm.

3.1. LDA

A topic model is a type of statistical model for discovering a set of topics that describe a collection of documents. The most common topic model currently in use is the LDA model proposed by Blei et al. (2003). The model generates automatic summaries of topics in terms of a discrete probability distribution over words for each topic, and further infers per-document discrete distributions over topics. The interaction between the observed documents and hidden topic structure is manifested in the probabilistic generative process associated with LDA. This generative process can be thought of as a random process that is assumed to have produced the observed documents. Let M, N, K, and V be the number of documents in a corpus, the number of words in a document, the number of topics, and the vocabulary size, respectively. The notations $Dirichlet(\cdot)$ and $Multinomial(\cdot)$ represent Dirichlet and multinomial distribution with parameter $(\cdot)$, respectively. The notation $\boldsymbol{\beta}_k$ is the V-dimensional word distribution for topic k, and $\boldsymbol{\theta}_d$ is the K-dimensional topic proportion for document d. The notations $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$ represent the hyperparameters of the corresponding Dirichlet distributions. The graphical representation of LDA is shown in Figure 1, and the corresponding generative process is shown below:

(1) For each topic $k \in \{1, \ldots, K\}$,
  (a) draw a distribution over vocabulary words $\boldsymbol{\beta}_k \sim Dirichlet(\boldsymbol{\eta})$.
(2) For each document d,
  (a) draw a vector of topic proportions $\boldsymbol{\theta}_d \sim Dirichlet(\boldsymbol{\alpha})$.
  (b) For each word $w_{d,n}$ in document d,
    (i) draw a topic assignment $\mathbf{z}_{d,n} \sim Multinomial(\boldsymbol{\theta}_d)$;
    (ii) draw a word $w_{d,n} \sim Multinomial(\boldsymbol{\beta}_{\mathbf{z}_{d,n}})$.

Figure 1 Graphical Model of LDA



3.2. Sent-LDA

The original LDA model is based on the bag-of-words assumption, which states that the order of words in a document does not matter. But this assumption is clearly unrealistic in our case since we observe (and will validate later) that each sentence (risk factor) in a document is only regarding one risk type (topic) in most cases. Intuitively, sentence boundaries convey the information about what words should be grouped into the same topic, and this information should be able to enhance the model by steering it toward more meaningful topics for representing risk types. In contrast, under the bag-of-words assumption, the boundaries between sentences will be ignored and the words in a sentence will be sampled independently from each other. This might result in scenarios where each word in a sentence is sampled from a different topic, severely violating our observation. To alleviate this issue, we propose to take the boundaries between sentences into account and assume that all words in a sentence are sampled from the same topic. This relaxes the bag-of-words assumption in the sense that the words in different sentences are no longer interchangeable and the sampling of the words in the same sentence are dependent on each other.

It is worth mentioning that some recently proposed methods do exploit sentence structures to enhance the LDA model. Distinct from our proposed "one topic per sentence" assumption, all these methods allow each sentence to include multiple topics, and use various means to incorporate sentence structure. The most straightforward method is to treat each sentence as a document and apply the LDA model on the collection of sentences rather than documents. Despite its simplicity, this method, called local-LDA (Brody and Elhadad 2010), has been demonstrated to be effective in discovering meaningful topics while summarizing consumer reviews. Another variant (Titov and McDonald 2008, Chang and Chien 2009, Wang et al. 2009, Du et al. 2010, Lin et al. 2011) models the sentence-wide topic proportion in addition to the document-wide topic proportion in original LDA model. In particular, the topics of words in a sentence are allowed to be sampled from either document-wide or sentence-wide topic proportions. These sentence-wide topic proportions are used to model the emphasis of each sentence and can be varied across sentences in a document. More recently, Lu et al. (2011) compared several aforementioned methods (Titov and McDonald 2008, Du et al. 2010, Brody and Elhadad 2010) for the task of labeling sentences with ratable aspects (i.e., topics) in product reviews, and found that local-LDA (Brody and Elhadad 2010) performs best despite its simplicity. Thus, we will choose local-LDA as one of our benchmark methods in our evaluation later.

We now describe our proposed model, called sent-LDA, which makes the "one topic per sentence" assumption. Using the same notations for LDA and letting S be the number of sentences in a document, the generative process of our sent-LDA is changed to the following:

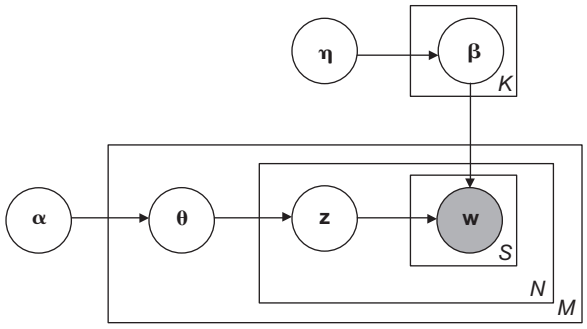
(1) For each topic $k \in \{1, \ldots, K\}$,
  (a) draw a distribution over vocabulary words $\boldsymbol{\beta}_k \sim Dirichlet(\boldsymbol{\eta})$.
(2) For each document d,
  (a) draw a vector of topic proportions $\boldsymbol{\theta}_d \sim Dirichlet(\boldsymbol{\alpha})$.
  (b) For each sentence s in document d,
    (i) draw a topic assignment $\mathbf{z}_{d,s} \sim Multinomial(\boldsymbol{\theta}_d)$.
    (ii) For each word $w_{d,s,n}$ in sentence s,
      (ii-a) draw a word $w_{d,s,n} \sim Multinomial(\boldsymbol{\beta}_{\mathbf{z}_{d,s}})$.

Figure 2 presents the graphical representation of our sent-LDA model, which adds a sentence layer in the original hierarchy of LDA in Figure 1.

3.2.1. Selection of Learning Algorithm. The key inferential problem we need to solve in order to use LDA and our sent-LDA topic model is that of computing the posterior distribution of the hidden variables $\boldsymbol{\theta}$ (topic proportions) and $\mathbf{z}$ (topic assignments) given the model parameters and the observed documents $\mathbf{w}$:

$$p(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\mathbf{w})}. \quad (1)$$

Figure 2 Graphical Model of Sent-LDA



Unfortunately, this distribution is intractable to compute in general (Blei et al. 2003). To tackle this problem, topic model learning algorithms attempt to approximate Equation (1) by forming an alternative distribution over the latent topic structure that is adapted to be close to the true posterior. The learning algorithms generally fall into two categories: *sampling-based algorithms* and *variational algorithms*. Sampling-based algorithms attempt to collect samples from the posterior to approximate it with an empirical distribution. Rather than approximating the posterior with samples, variational methods posit a parametrized family of distributions over the hidden structure and then solve an optimization problem to find the member of that family that is closest to the posterior. *Collapsed Gibbs sampling* (Griffiths and Steyvers 2004) and *variational EM algorithm* (Blei et al. 2003) are the most commonly used sampling-based and variational methods, respectively.

There are many discussions on the advantages and disadvantages of these two types of learning algorithms, and some previous studies (Teh et al. 2007, Asuncion et al. 2009, Wallach et al. 2009, Zhai et al. 2012) have attempted to compare their performance. However, the findings are mixed. Following Blei and Jordan (2006), we resort to the empirical experiments for comparing the different learning algorithms in our context. Note that Jo and Oh (2011) have proposed a model that is equivalent to our sent-LDA model assumption, but used the collapsed Gibbs sampling method for learning. As we will demonstrate later, this sent-LDA-CGS model performs even worse than the original LDA model. In contrast, we design the variational EM learning algorithm for the sent-LDA model, which performs best compared with benchmark methods.

3.2.2. Approximate Inference. In posterior inference, we compute the conditional distribution of latent variables given a set of observed documents. This conditional distribution for our sent-LDA is the same as that of the LDA shown in Equation (1). However, the interpretation of the vector $\mathbf{z}$ is changed. Specifically, since we only draw the topic assignment for each sentence (the words in a sentence share the same topic assignment) rather than each word, $\mathbf{z}$ is now the vector of topic assignments for sentences rather than words in a document.

Variational methods consider a simple family of distributions over the latent variables, indexed by free variational parameters, and try to find the setting of those parameters that minimizes the Kullback–Leibler divergence to the true posterior. In the sent-LDA model, the latent variables are the per-document topic proportion $\boldsymbol{\theta}$ and the per-sentence topic assignment $\mathbf{z}$. Similar to variational methods for LDA, we use the variational distribution

$$q(\boldsymbol{\theta}, \mathbf{z} \mid \boldsymbol{\gamma}, \boldsymbol{\phi}) = q(\boldsymbol{\theta} \mid \boldsymbol{\gamma}) \prod_{s=1}^{S} q(\mathbf{z}_s \mid \boldsymbol{\phi}_s)$$

as a surrogate for the posterior distribution in Equation (1).

We now describe how to set the variational parameter $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ via an optimization procedure. We bound the log-likelihood of a document using Jensen's inequality. By omitting the variational parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$, we have

$$\begin{aligned}
\log p(w \mid \alpha, \beta) &= \log \int \sum_z p(\theta, z, w \mid \alpha, \beta)\, d\theta \\
&= \log \int \sum_z \frac{p(\theta, z, w \mid \alpha, \beta) q(\theta, z)}{q(\theta, z)}\, d\theta \\
&\geq \int \sum_z q(\theta, z) \log p(\theta, z, w \mid \alpha, \beta)\, d\theta \\
&\quad - \int \sum_z q(\theta, z) \log q(\theta, z)\, d\theta \\
&= E_q[\log p(\theta, z, w \mid \alpha, \beta)] - E_q[\log q(\theta, z)] \\
&= L(\boldsymbol{\gamma}, \boldsymbol{\phi}; \alpha, \beta).
\end{aligned}$$

By expanding the lower bound L using the factorization of p and q, we have

$$\begin{aligned}
&L(\boldsymbol{\gamma}, \boldsymbol{\phi}; \alpha, \beta) \\
&= E_q[\log p(\theta \mid \alpha)] + E_q[\log p(z \mid \theta)] + E_q[\log p(w \mid z, \beta)] \\
&\quad - E_q[\log q(\theta)] - E_q[\log q(z)] \\
&= \log \Gamma\left(\sum_{j=1}^{K} \alpha_j\right) - \sum_{i=1}^{K} \Gamma(\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1)\left(\psi(\gamma_i) - \psi\left(\sum_{j=1}^{K} \gamma_j\right)\right) \\
&\quad + \sum_{s=1}^{S} \sum_{i=1}^{K} \phi_{si}\left(\psi(\gamma_i) - \psi\left(\sum_{j=1}^{K} \gamma_j\right)\right) \\
&\quad + \sum_{s=1}^{S} \sum_{i=1}^{K} \phi_{si} \sum_{n=1}^{N_s} \sum_{j=1}^{V} w_n^j \log \beta_{ij} \\
&\quad - \log \Gamma\left(\sum_{j=1}^{K} \gamma_j\right) + \sum_{i=1}^{K} \log \Gamma(\gamma_i) - \sum_{i=1}^{K} (\gamma_i - 1)\left(\psi(\gamma_i) - \psi\left(\sum_{j=1}^{K} \gamma_j\right)\right) \\
&\quad - \sum_{s=1}^{S} \sum_{i=1}^{K} \phi_{si} \log \phi_{si},
\end{aligned}$$

where ψ is the first derivative of the $\log \Gamma$ function, N_s is the number of words in sentence s, and w_n^j equals 1 if word w_n is the jth word in the vocabulary and 0 otherwise. Each line on the right-hand side of the second equal sign corresponds to each term on the right-hand side of the first equal sign. Note that the difference between our expanded lower bound and that of the LDA in Blei et al. (2003) lies in the second, third, and fifth terms because of the additional sentence layer.

Maximizing lower bound $L(\gamma, \phi; \alpha, \beta)$ with respect to the variational parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$, we obtain the following update equations:

$$\phi_{si} \propto \left(\prod_{n=1}^{N_s} \beta_{iw_n}\right) \exp\left(\psi(\gamma_i) - \psi\left(\sum_{j=1}^{K} \gamma_j\right)\right)$$

$$\gamma_i = \alpha_i + \sum_{s=1}^{S} \phi_{si},$$

where ϕ_{si} is the probability that sentence s is generated by topic i, and γ_i is the ith component of posterior Dirichlet parameter.

3.2.3. Parameter Estimation. Given a corpus of documents, we aim to find parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that maximize the log-likelihood of the observed data. To achieve this objective, we use a variational EM procedure as in Blei et al. (2003). In the E-step, we find the optimizing values of the variational parameters for each document. This is done as described in the previous inference subsection. In the M-step, we find the maximum likelihood estimates of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ using expected sufficient statistics computed in the E-step. These two steps are repeated until the lower bound on log-likelihood converges.

By fixing the values of variational parameters and maximizing the lower bound of likelihood with respect to the model parameters, we obtain the M-step update for the multinomial parameter $\boldsymbol{\beta}$:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \sum_{i=1}^{K} \phi_{dsni} w_{dsn}^{j},$$

where β_{ij} is the probability that the jth word is generated by topic i, all words w_{dsn} in sentence s share the same ϕ_{si}, and w_{dsn}^{j} equals 1 if word w_{dsn} in sentence s of document d is the jth word in the vocabulary and 0 otherwise.

For the Dirichlet parameter $\boldsymbol{\alpha}$, we cannot derive the closed form of its M-step update. We use the Newton–Raphson algorithm described in Blei et al. (2003) to find its optimal value.

4. Experiments

In this section, we conduct comprehensive experiments to evaluate our proposed method. We first describe the data preparation and benchmark methods. We then compare our proposed method with the competing unsupervised methods in terms of predictive power, clustering quality, and the quality of discovered information. For a more convincing validation, we further compare our method with the supervised method. Finally, we discuss the choosing of the number of topics and the computational complexity of our algorithm.

4.1. Data Preparation

To prepare our data set, we extract the textual risk factors in section 1A (a newly created section since 2005) of each 10-K form as a document. The 10-K forms across five years from 2006 to 2010 are collected from EDGAR databases on the SEC's website.[3] For each risk factor, we only retain the summary heading as shown in Table 1. Because of the inconsistent file format (e.g., TXT or HTML) and form layout (e.g., headings are highlighted using different fonts or capitalized letters), it is quite challenging to automatically extract these risk factors from 10-K forms. To deal with these issues, we parse the HTML files into a tree structure and then scrape needed information using predefined heuristic rules. For the TXT files, we create a set of heuristic rules, taking into account the section title, section position, section length, and so on, to retain the needed risk factors. Since our heuristics depend on the structure of the form text, we might end up with some "noise," i.e., misextracted content. As we will report later in this section, we manually analyze the accumulated text, and find that the relative amount of such noise is quite low, indicating good quality of extraction. Through this process, we obtain our data set consisting of 14,799 documents and 322,287 sentences (21.78 sentences per document on average) of risk factor disclosures in section 1A of 10-K forms.

To compare our unsupervised method with supervised methods, we have to construct a training set for learning supervised models. The construction of training sets consists of two steps. First, we have to predefine a set of risk types (categories) and create a coding scheme accordingly. To this end, we directly adopt the taxonomy of risk types proposed by Huang and Li (2011), who are experts in financial accounting and have defined 25 risk types by reading hundreds of annual reports. In addition to those 25 risk types, we add the other two categories for coding, namely "other risk types" and "not a risk type." "Other risk types" is added since we found that there are many risk factor sentences that do not belong to any of those 25 risk types; "not a risk type" is added since there are some misextracted content as aforementioned. Second, we have to select a subset of risk factors (sentences) that are representative of the corpus. Since random sampling is most appropriate for obtaining a representative sample (Grimmer and Stewart 2013), we randomly sample 3,000 out of 322,287 sentences for labeling.

We recruited four graduate students to label the sampled risk factor sentences. These students are native English speakers and have taken courses in financial accounting. Each student labels 1,500 out of 3,000 risk factors and each risk factor is labeled by two students. Before labeling, they are briefed on the definition and

[3] http://www.sec.gov/edgar/searchedgar/ftpusers.htm (accessed March 22, 2014).

trained on a number of real labeled examples of each risk type. As an incentive, each student was paid $50. To measure the inter-rater agreement when labeling the training set, we calculate Cohen's kappa and the corresponding maximum kappa. The maximum kappa is usually reported to assist the interpretation of kappa value as suggested by Sim and Wright (2005). In our case, Cohen's kappa value is 0.5679 (with Max Cohen's kappa value of 0.8612), indicating a moderate strength of inter-rater agreement according to Sim and Wright (2005). To ensure the consistency, we only retain the risk factor sentences whose labels are agreed upon by all annotators. This leads to a set of 1,842 examples. After removing examples labeled with "other risk types" and "not a risk type," we obtain a training set of 1,327 examples.

As mentioned previously, because of the heuristic nature of our extraction procedure, we end up with some misextracted text. During the manual labeling procedure, these are awarded the label of "not a risk type." At the end of the labeling task, we counted the number of such misextracted sentences, and found that only 17 of the 1,842 extracted sentences (0.92%) possessed a "not a risk type" label, indicating the robustness of our extraction heuristics.

To validate our fundamental assumption that each risk factor (sentence) only discusses one topic, we additionally require that each annotator records all risk factors that might belong to multiple labels. It turns out that there are only 11 such risk factors, making up 0.83% (11/1,327) of the total. Clearly, in an overwhelming majority of cases, there is a one-to-one mapping between sentences and topics. This validates the key assumption of our proposed model.

4.2. Benchmark Methods

To compare the performance of our proposed method with other unsupervised learning methods, we adopt two benchmark models: the original LDA model and the local-LDA model. The local-LDA (Brody and Elhadad 2010) directly applies LDA on the collection of sentences rather than documents. To examine the effect of learning algorithms, we learn each model with the two learning algorithms discussed previously, namely, the variational EM (VEM) and the collapsed Gibbs sampling (CGS) algorithm. By pairing each model with each learning algorithm, we obtain six methods denoted as *sent-LDA-VEM*, *sent-LDA-CGS*, *local-LDA-VEM*, *local-LDA-CGS*, *LDA-VEM*, and *LDA-CGS*, respectively. Sent-LDA-VEM is our proposed method and the others are baselines. It is worth noting that sent-LDA model assigns topics at the sentence level and LDA and local-LDA assign topics at the word level. To use LDA and local-LDA for our task, it requires an additional step to calculate the sentence-level topic assignment based on the inferred topics of words in the sentence.

To perform fair comparisons, we use the same parameter settings for all methods. Specifically, for the variational EM learning algorithms, the maximum number of EM iterations is 1,000, and the likelihood convergence criteria is 1×10^{-5}. For the collapsed Gibbs sampling, we set the hyperparameters as suggested by Griffiths and Steyvers (2004)—α is set to $50/k$ where k is the number of topics, and η is set to 0.1. The number of iterations is set to 2,000.

To compare the performance of our proposed method with supervised learning methods, we implement the state-of-the-art categorical K-nearest neighbors (CKNN) algorithm (Huang and Li 2011) for categorizing textural risk factor disclosures. Since we assume that each risk factor is only regarding one risk type, we simply classify each risk factor with the most probable risk type generated by the CKNN algorithm.

4.3. Predictive Power

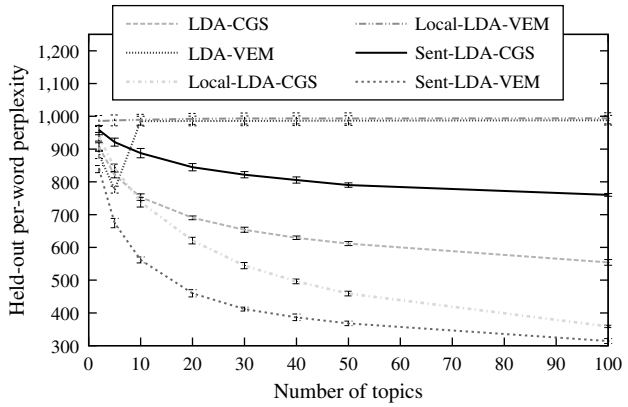
The most typical evaluation of topic models involves measuring how well a model performs when predicting unobserved documents. Specifically, when estimating the probability of unseen held-out document given a set of training documents, a "good" model should give rise to a higher probability of held-out documents. To measure the predictive power of competing models, we use a metric, called perplexity, that is conventional in language modeling (Azzopardi et al. 2003). The perplexity can be understood as the predicted number of equally likely words for a word position on average, and is a monotonically decreasing function of the log-likelihood. Thus, a lower perplexity over a held-out document is equivalent to a higher log-likelihood, which indicates better predictive performance. Formally, for a test set D_{test} of M documents, the per-word perplexity is defined as

$$perplexity(D_{\text{test}}) = \exp\left(-\sum_{d=1}^{M} \log p(w_d) \Big/ \sum_{d=1}^{M} N_d\right),$$

where N_d is the number of words in document d. To ensure consistency of evaluation across models when computing perplexity, we follow the approximation of Teh et al. (2008) of the predictive likelihood $p(w_d \mid D_{\text{train}})$ using $p(w_d \mid D_{\text{train}}) \approx p(w_d \mid \hat{\theta}_d)$, where $\hat{\theta}_d$ is a point estimate of the posterior topic proportions for document d.

Figure 3 shows the predictive power of each model in terms of the held-out per-word perplexity by varying the number of topics (the deviations are shown as error bars). This figure is obtained via tenfold cross-validation as in Blei and Lafferty (2007). Specifically, we first divide the data into 10 folds. For each fold i and each model, we fit the model to the data that are not in fold i and then use the fitted model to do inference for the data in fold i. Then the held-out metrics (e.g., per-word perplexity) in each fold can be

Figure 3 **Held-Out per-Word Perplexity as a Function of the Number of Topics**



computed. As can be seen, our proposed sent-LDA-VEM performs best and achieves the lowest perplexity for all the number of topics. In terms of the effects of learning algorithms, it is interesting to observe that collapsed Gibbs sampling leads to better performance than variational EM for LDA and local-LDA, but results in worse performance for sent-LDA. In terms of the effects of the number of topics, the perplexity of all methods monotonically decreases with the increase of the number of topics, but tends to converge to a fixed value eventually. When the number of topics is larger than 30, the perplexity tends to be steady.

Table 2 **Comparison of Models with 30 Topics in Terms of Held-Out per-Word Perplexity**

	LDA-CGS	LDA-VEM	Local-LDA-CGS	Local-LDA-VEM	Sent-LDA-CGS	Sent-LDA-VEM
Mean	632.91	967.11	524.47	973.78	804.33	389.04
Std. dev.	(±8.26)	(±16.84)	(±8.74)	(±18.13)	(±10.17)	(±10.45)
p-value	0.0000	0.0000	0.0000	0.0000	0.0000	—

To test the significance of performance difference between benchmark methods and our proposed method, we present the perplexity of all methods with 30 topics and conduct the paired *t*-tests with our proposed sent-LDA-VEM as shown in Table 2. As can be seen, our proposed sent-LDA-VEM significantly outperforms all the baseline methods at a 1% significance level.

To examine the efficiency of different combinations of models and learning algorithms, we plot the per-word perplexity and the empirical log-likelihood of the training data during model learning. We plot the model performance as a function of the number of iterations of the learning algorithm in Figure 4. We also plot the model performance as a function of running time in Figure 5 since VEM algorithms usually need

Figure 4 **Per-Word Perplexity and Empirical Log-Likelihood as a Function of the Number of Iterations**

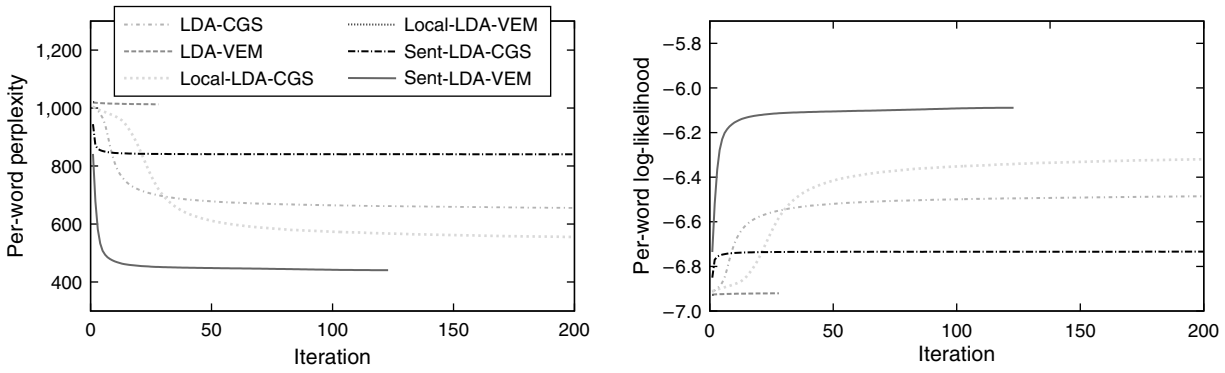


Figure 5 **Per-Word Perplexity and Empirical Log-Likelihood as a Function of Time**

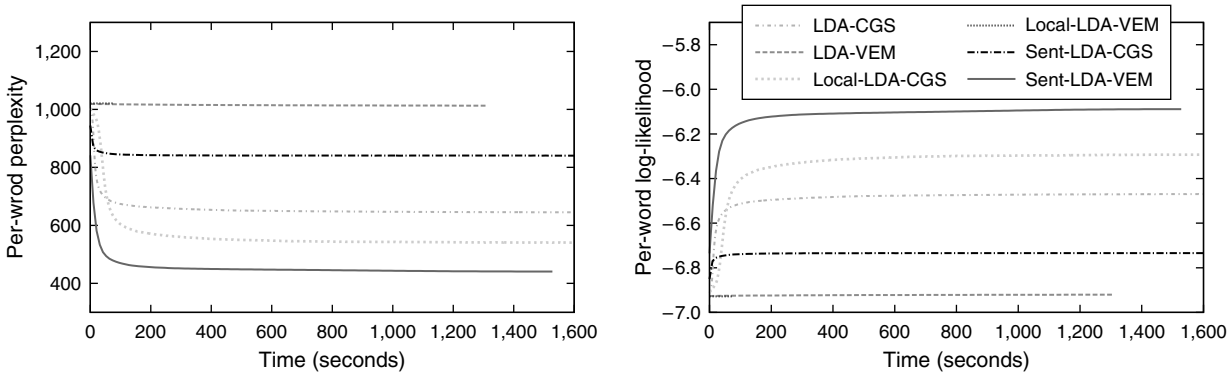


Table 3 Comparison of Models with 30 Topics in Terms of Silhouette Coefficients

	LDA-CGS	LDA-VEM	Local-LDA-CGS	Local-LDA-VEM	Sent-LDA-CGS	Sent-LDA-VEM
Mean	−0.06358	−0.08186	−0.03284	−0.12421	−0.02975	−0.02932
Std. dev.	(±0.01066)	(±0.01847)	(±0.00786)	(±0.05267)	(±0.00673)	(±0.01064)
p-value	0.0000	0.0000	0.4112	0.0000	0.9152	—

dozens of iterations to converge, and CGS algorithms require thousands of shorter iterations (Zhai et al. 2012). As shown in Figures 4 and 5, all the algorithms tend to converge quickly (within 50 iterations and 100 seconds). When converged, our sent-LDA-VEM model achieves the lowest per-word perplexity and highest log-likelihood. Note that VEM algorithms will stop when the convergence criteria are met, but it is difficult to determine the convergence criteria for CGS algorithms (Zhai et al. 2012).

4.4. Cluster Quality

Cluster quality refers to the extent to which intracluster similarities outdistance intercluster similarities. To measure the cluster quality of competing models, we use the *silhouette coefficient* metric (Rousseeuw 1987). For a given point i, the silhouette of i is defined as $s(i) = (b(i) - a(i))/\max\{a(i), b(i)\}$, where $a(i)$ is the average distance between point i to all other points in the same cluster, and $b(i)$ is the average distance between point i to the points in the nearest cluster, which i is not a member. The silhouette of a data sample is the average of silhouettes of all points. The silhouette is bounded between −1 (for bad clustering) and +1 (for highly dense clustering).

Table 3 shows the silhouette coefficient for all methods with 30 topics. Here, we treat each sentence as a data point, and the most probable topic as the corresponding cluster. For calculating, we use Euclidean distance. Test statistics are computed via tenfold cross-validation in the same way as that in Table 2. As can be seen, our sent-LDA-VEM performs best among all models. It significantly outperforms LDA-CGS, LDA-VEM, and local-LDA-VEM, but performs equally well as local-LDA-CGS, and sent-LDA-CGS.

It should be noted that silhouette coefficients assume hard clustering, whereas topic models actually perform soft clustering where each object might belong to multiple clusters (topics) with different probabilities. Thus, it is not as suitable as other metrics like perplexity or those that will be introduced later. But its advantage is that it does not require ground-truth data, which might be expensive to obtain for unsupervised learning methods.

4.5. Quality of Discovered Information

The previously reported objective evaluation metrics (i.e., perplexity and silhouette coefficient) are essential for understanding model performance and are commonly used in the computer science community. However, it is more important to evaluate the quality of the discovered information if the goal is to use unsupervised topic models for social science research. To this end, we next investigate the quality of the discovered information.

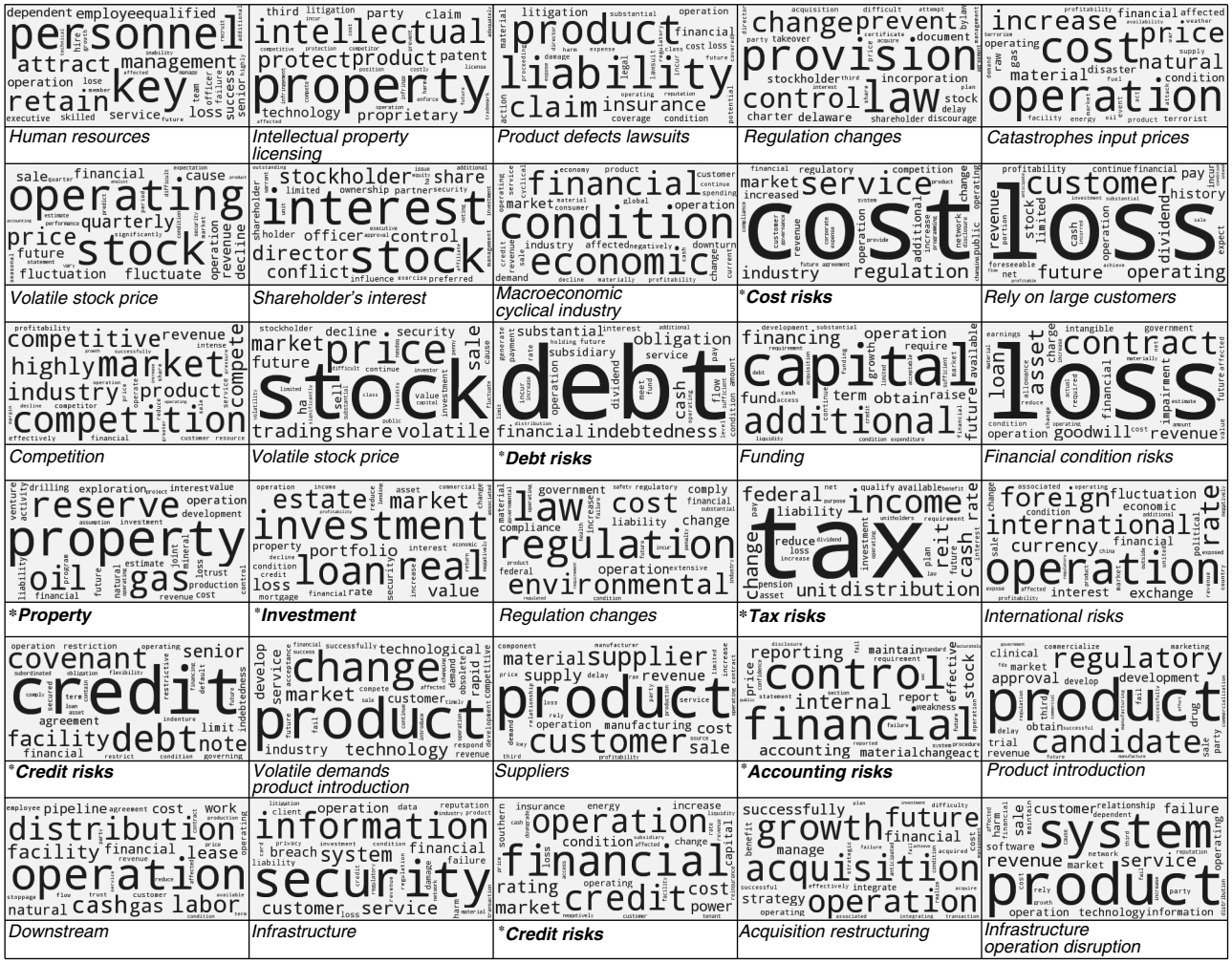
4.5.1. Labeling Topics. Before using or validating the topics learned by topic models, the topics need to be labeled so that we could determine what each topic measures. There exist some automatic labeling methods (Mei et al. 2007), but these are not suitable in cases where such labeling requires domain knowledge (financial knowledge in our case). Actually, in most topic model research, it is customary to manually label topics ensuring high labeling quality (Chang et al. 2009). We thus design a manual labeling procedure that makes use of human experts' domain knowledge. In particular, we first adopt 25 risk types defined by Huang and Li (2011) as the set of candidate labels, and attempt to map topics to these 25 labels as well as possible. For those topics that cannot be mapped to any of those 25 labels, we mark them with "other risk types," and label them later with new meaningful label names suggested by domain experts.

To execute this procedure, we recruit two human annotators to label the topics learned by LDA-CGS, local-LDA-CGS, and sent-LDA-VEM. The number of topics for each model is set to 30, and the most effective learning algorithm is chosen for each model. To ensure consistency, the annotators are selected from four human "labelers" chosen for creating the training set. They first perform the mapping on their own. Table 4 reports the inter-rater agreement for mapping the topics of each model. As can be seen, the annotators achieve almost perfect agreement ($\kappa = 0.8400$) for sent-LDA-VEM model, substantial agreement ($\kappa = 0.6296$) for local-LDA-CGS model, and moderate agreement ($\kappa = 0.4958$) for LDA-CGS model. This observation demonstrates the superiority of our sent-LDA-VEM model since good topics should be more representative for risk types and thus easier to be labeled.

Table 4 Inter-Rater Agreement for Labeling Topics

	Cohen's kappa	Max Cohen's kappa
LDA-CGS	0.4958	0.6639
Local-LDA-CGS	0.6296	0.7407
Sent-LDA-VEM	0.8400	0.8400

Figure 6 Topic Labeling



Note. Topics learned by our sent-LDA-VEM are visualized using word clouds. Risk type labels defined in Huang and Li (2011) are italicized, and new risk type labels are italicized, bolded, and preceded by an asterisk (*). *Topic*1 to *topic*30 are displayed from left to right, top to bottom.

After the independent mapping, the annotators then get together to achieve consensus, and then decide the labels for topics marked with "other risk types." Figure 6 presents the labeled topics learned by our sent-LDA-VEM model with 30 topics. Each topic is visualized using word clouds, where the font size corresponds to the probability of the word occurring in the topic.

Referring back to the debate regarding the competition between unsupervised and supervised learning methods mentioned in §1, we take Figure 6 as an example for illustrating how our unsupervised topic model can be used for suggesting a classification scheme for supervised methods when the taxonomy is unclear. As admitted by Huang and Li (2011), their taxonomy of 25 risk types is defined based on their subjective judgment, and some important risk factor types may be left out. This is further confirmed by the fact that we find additionally 498 examples (accounting for 37.6% (498/1,327) of total training examples) labeled with "other risk types" that cannot be categorized into any of their 25 risk types when constructing our training set described in §4.1. As shown in Figure 6, our learned topics via the unsupervised topic model could find all those 25 risk types, although some highly related types are merged together. More importantly, we find some additional risk types including "cost risks," "debt risks," "property risks," "investment risks," "tax risks," "credit risks," and "accounting risks." We have verified the joint significance of these newly discovered risk types that will appear in our empirical study later. Thus, if we resort to our unsupervised method when defining the taxonomy for supervised learning method, we could reduce the risk of missing some important risk types.

4.5.2. Validating Topics. To validate the quality of topics, most topic modeling works only provide qualitative assessments of inferred topics (as lists of ranked keywords) and simply assert that topics are

semantically meaningful. Chang et al. (2009) emphasize that not measuring the internal representation (latent topics) of topic models is at odds with their presentation and development. To address this issue, Chang et al. (2009) and Grimmer and King (2011) have recently developed some measures based on elicited judgment by subject experts. In the following, we employ a number of such measures to quantitatively validate our inferred topics.

Semantic Validation. The semantic coherence of topics is perhaps the most important indicator of the quality of topics. Different from standard clustering methods, topic models yield a set of keyword lists (or more formally, multinomial distributions over words) for each cluster (topic). Semantic coherence of a topic refers to how well the topic matches a human concept based on its keyword list.

To quantitatively measure the coherence of topics, we adopt the *word intrusion* task designed by Chang et al. (2009). In the word intrusion task, the subject is presented with six randomly ordered words. The task of the subjects is to find the word that is out of place or does not belong with others, i.e., the *intruder*. When the set of words minus the intruder makes sense together, the subjects should easily identify the intruder. For example, for a set of words {*dog*, *cat*, *horse*, *apple*, *pig*, *cow*}, the word "*apple*" is easily identified as the intruder since all the other words refer to animals. In contrast, for a set of words {*car*, *teacher*, *cat*, *pig*, *bike*, *cup*}, which lacks coherence, it is difficult to identify the intruder.

To construct the set to present to the subjects, we follow the procedures by Chang et al. (2009). First, we randomly select a topic inferred by a model, and select the five most probable words from that topic. In addition to these words, an intruder word is selected at random from a pool of words with low probability in the current topic (to reduce the possibility that the intruder from the same semantic group) but high probability in some other topic (to ensure that the intruder is not rejected solely due to rarity). All six words are then shuffled and presented to the subjects. The *model precision* MP_m^k of the kth topic inferred by model m in word intrusion task is defined as the fraction of subjects agreeing with model

$$MP_m^k = \frac{1}{S}\sum_s \mathbb{1}(i_{k,s}^m = w_k^m),$$

where $i_{k,s}^m$ is the intruder word selected by subject s among S subjects, w_k^m is the true intruder word, and $\mathbb{1}(\cdot)$ is an indicator function that equals 1 if $(\cdot)$ is true and 0 otherwise. The model precision MP_m of model m is simply the average of corresponding MP_m^k over topics.

Figure 7 Model Precision for Word Intrusion Task

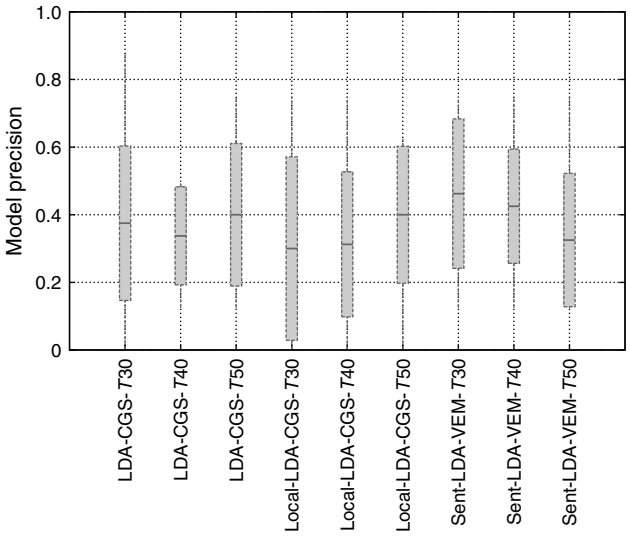
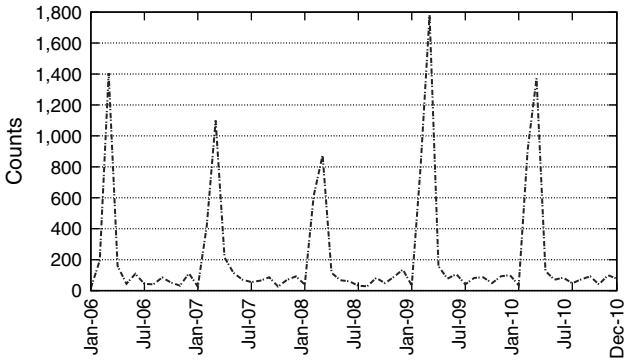


Figure 7 shows boxplots of model precision of three models (LDA-CGS, local-LDA-CGS, and sent-LDA-VEM) with different numbers of topics ($T30$, $T40$, $T50$). The most effective learning algorithm is chosen for each model, and the number of topics is set to 30, 40, 50 because the perplexity in Figure 3 begins to converge in the range [30, 50]. We observe that the model precision will be affected by the number of topics. Specifically, our sent-LDA-VEM performs better than the other two models when the number of topics is set to 30 and 40, but worse when the number of topics is set to 50. Overall, our proposed sent-LDA-VEM model with 30 topics performs best. It is interesting to observe that the model performance in the word intrusion task is not consistent with the performance in terms of the predictive power as shown in Figure 3. This means that a model with more predictive power does not necessarily ensure a higher quality of inferred topics in terms of semantic coherence. This observation is consistent with that in Chang et al. (2009), and sheds some light on the model selection and model parameter (i.e., number of topics) settings, which we will discuss later.

Predictive Validation. Quinn et al. (2010) and Grimmer (2010) argue that if topics are valid, external events should explain sudden increases in attention to a topic. Following these works, we perform a similar predictive validation using external events. Figure 8 plots the number of risk factors (sentences) released each month about "macroeconomics risks" inferred by our sent-LDA-VEM model as shown in Figure 6. The count of risk factors double around the year 2009, probably because of the financial crisis. This shows that the external event (financial crisis) predicts the spikes in attention in textual corporate risk disclosures.

Topic Assignment Validation. Recall that the output of topic models has two components: one is the makeup

Figure 8 Predictive Validation of the Topic "Macroeconomics Risks"



of topics, which is validated via the word intrusion task previously; the other one is the topic assignment for each word (LDA and local-LDA) or sentence (sent-LDA) of documents. Consequently, it is also important to test whether the topic assignments make sense.

To provide an intuitive example of topic assignments for sentences (risk factors) by using our proposed sent-LDA-VEM model, we conduct a case analysis of Apple Inc's 10-K from the year 2006 and present some examples of topic assignments in Table 5. The number of topics of sent-LDA-VEM is set to 30 and the word cloud of each topic label can be found in Figure 6. One observation is that each risk factor indeed corresponds to only one risk type (topic), which again confirms our key assumption that each sentence (risk factor) can be only assigned for one topic. Another observation is that the assigned topic label well categorizes the corresponding risk factors.

Here, we only provide a qualitative case analysis of topic assignments for our sent-LDA-VEM model. A more reliable quantitative validation of topic assignments will be presented later when comparing our model with the supervised ones.

4.6. Comparison with Supervised Learning Method

Although we have reported several evaluations of our proposed unsupervised topic model, one might still be skeptical regarding its performance because of the lack of ground-truth data. For this reason, there is, typically, much confidence in the evaluation results of supervised methods since the availability of ground-truth data (training data) is a prerequisite for learning and thus can be utilized when validating. To alleviate this skepticism and conduct an equally valid evaluation of our unsupervised method, we construct a training set as in §4.1 and use it as the ground-truth data for the task of risk factor classification.

To use the output of our unsupervised topic model for risk factor classification defined in Huang and Li (2011), we first need to map the inferred topics to their predefined 25 risk types as shown in Figure 6. After mapping, we can easily classify each risk factor based on the assigned topics. To compare the performance of unsupervised methods with the supervised ones, we implemented the state-of-the-art categorical K-nearest neighbor algorithm proposed by Huang and Li (2011) for risk factor classification.

Table 5 Some Examples of Topic Assignments for Risk Factors Disclosed by Apple Inc. in 2006

[Topic label] risk factors
[T1: human resources risks] The company's success depends largely on its ability to attract and retain key personnel.
[T2: intellectual property risks] The company's business relies on access to patents and intellectual property obtained from third parties, and the company's future results could be adversely affected if it is alleged or found to have infringed on the intellectual property rights of others.
[T3: potential/ongoing lawsuits] Unfavorable results of legal proceedings could adversely affect the company's results of operations.
[T5: catastrophes] War, terrorism, public health issues, and other circumstances could disrupt supply, delivery, or demand of products, which could negatively affect the company's operations and performance.
[T7: macroeconomic risks] Economic conditions and political events could adversely affect the demand for the company's products and the financial health of its suppliers, distributors, and resellers.
[T12: volatile stock price] The company's stock price may be volatile.
[T20: international risks] The company's business is subject to the risks of international operations.
[T22: new product introduction] The company must successfully manage frequent product introductions and transitions to remain competitive and effectively stimulate customer demand.
[T23: suppliers risks] Future operating results are dependent upon the company's ability to obtain a sufficient supply of components, including microprocessors, some of which are in short supply or available only from limited sources.
[T27: infrastructure] Failure of information technology systems and breaches in the security of data upon which the company relies could adversely affect the company's future operating results.

Table 6 shows the fivefold cross-validation classification accuracy of the supervised CKNN method with different numbers of neighbors k, and our sent-LDA-VEM with number of topics $T = 30$. As can be seen, CKNN performs best when $k = 5$; thus, we conduct the t-test for all the other methods paired with it. The performance of our proposed sent-LDA-VEM model is not significantly (p-value $= 0.3720$) different from the best supervised method CKNN ($k = 5$), which indicates that it performs equally well as the state-of-the-art supervised learning methods.

At this point, it is useful to reiterate the Grimmer and Stewart (2013) observation, that the validation of unsupervised methods in an supervised way *does not* obviate the need for unsupervised methods. This kind of validation is possible only after the unsupervised method suggests a classification scheme and provides one direct test to ensure that the output of an unsupervised method is just as valid, reliable, and useful as the supervised methods.

Table 6 Classification Accuracy by Fivefold Cross-Validation

	CKNN $k=2$	CKNN $k=5$	CKNN $k=10$	CKNN $k=15$	CKNN $k=20$	Sent-LDA-VEM $T=30$
Mean	0.8130	0.8362	0.8308	0.8233	0.8308	0.8255
Std. dev.	(±0.0155)	(±0.0216)	(±0.0294)	(±0.0170)	(±0.0145)	(±0.0134)
p-value	0.0869	—	0.7476	0.3237	0.6294	0.3720

4.7. Choosing the Number of Topics

Our proposed sent-LDA topic model is parametric, and the number of topics must be set beforehand. Determining the number of topics (clusters) is one of the most difficult questions in unsupervised learning. There are some methods that attempt to estimate the number of clusters automatically, but recent studies show that the estimated number of clusters are strongly model dependent (Wallach et al. 2010). It is also problematic to solely use fit statistics (such as perplexity or silhouette coefficient), as Chang et al. (2009) report that there is often a negative relationship between the best fitted model and the substantive information provided. Recently, Grimmer and Stewart (2013) notice this issue and argue that model selection should be recast as a problem of measuring *substantive fit* rather than *statistical fit*.

To determine the number of topics for our proposed model, we decide to take into account both statistical fit (in terms of perplexity as shown in Figure 3) and substantive fit (in terms of semantic coherence as shown in Figure 7). On one hand, choosing the number of topics based on perplexity relies on the assumption that the goal is to optimize the predictive power of the model. However, in the context where we seek to utilize unsupervised topic models for social science purposes, our goal is the revelation of substantively interesting information. To this end, we turn to substantive fit (semantic coherence). It turns out that measuring substantive fit (model precision in word intrusion task) needs human judgment, which is time consuming. Thus, we need to employ statistical fit to reduce the set of candidate models. Taking our proposed sent-LDA-VEM model as an example, we first choose 30, 40, and 50 to be the potential number of topics since its perplexity in Figure 3 tends to converge in the range [30, 50]. Then we compare the performance of our sent-LDA-VEM with 30, 40, and 50 topics as shown in Figure 7. Finally, we choose the number of topics to be 30 at which the model performance in word intrusion tasks is demonstrably the best.

4.8. Complexity Analysis

To demonstrate the efficiency of our proposed sent-LDA-VEM model, we use LDA-VEM model as a baseline and compare their computational complexities. We first analyze the computational complexity of the variational EM algorithm for LDA-VEM. The time complexity of E-step is $\mathcal{O}(MN^2K)$, where M is the number of documents in the corpus, N is the maximum document length in the corpus, and K is the number of topics. Actually, we only need to compute the posterior multinomial for the unique terms of each document for each iteration of variational inference, where the number of unique terms of a document must be slightly smaller than N. On the other hand, the time complexity of M-step is $\mathcal{O}(VK)$, where V is the vocabulary size. Thus, the main computational bottleneck of the variational EM algorithm for LDA is the E-step.

Next, we analyze the computational complexity of our derived variational EM algorithm for sent-LDA model. The time complexity is the same as that of LDA-VEM except that the time complexity of E-step for our model is $\mathcal{O}(MS^2K)$, where S is the number of sentences in a document. This is because we assume that all words in a sentence belong to the same topic and thus only need to compute the posterior multinomial for each sentence in a document. Thus, it is obvious that our sent-LDA-VEM model is more efficient than the LDA-VEM model since S will be definitely much smaller than N. In particular, in the same Linux system with dual 3.00 GHz CPU and 4.0 GB memory, to train a model with 30 topics against our data set, it takes 12.51 seconds on average for each iteration of our sent-LDA-VEM algorithm but 48.63 seconds on average for each iteration of the LDA-VEM algorithm.[4] More importantly, as shown in Figures 4 and 5, our sent-LDA-VEM method converges quickly to a much lower perplexity than the LDA-VEM and all the other benchmark models.

5. Empirical Study on Textual Risk Disclosures

In this section, we conduct an empirical study on the effect of textual risk disclosures on the postdisclosure risk perceptions of investors. We aim to demonstrate that our proposed sent-LDA-VEM model could facilitate the econometric analysis of textual disclosures by automatically converting them into manageable numeric variables. Besides, we aim to contribute to the literature on the effects of risk disclosures.

[4] We use the implementation of variational EM for LDA, available at http://www.cs.princeton.edu/~blei/lda-c/index.html (accessed March 22, 2014).

5.1. Research Hypotheses

Textual risk disclosures present users with firms' assessments about future contingencies, and they differ from other corporate disclosures in that they guide users about the *range* of future performance rather than the *level* of future performance (Kravet and Muslu 2011). Therefore, we hypothesize that the informative textual risk disclosures will change investors' risk perceptions, i.e., the range and confidence level in their predictions of future performance.

5.1.1. Proxy for Risk Perceptions. Stock return volatility is a prominent proxy for diverging investor opinions in the finance literature (Shalen 1993, Garfinkel 2009). Following Kravet and Muslu (2011), we predict higher daily stock return volatility during the first two months after the filings of disclosures than the last two months before the filings, reflecting the increased range and reduced confidence level in investors' prediction of future performance. On the other hand, if risk disclosures resolve known risk factors, investors will converge in their predictions and increase their confidence level, indicating a lower postdisclosure stock return volatility.

5.1.2. Measuring Textural Risk Disclosure. Different from previous studies, we use our proposed sent-LDA-VEM model (with 30 topics) to simultaneously discover and quantify risk types from textual risk disclosures in section 1A of 10-K forms. Risk types are aggregated at the document level, and each document (risk disclosures in a 10-K form) is a firm-year observation.

5.2. Research Design

5.2.1. Sample Description. Our initial sample includes all 10-K forms collected from 2006 to 2010. We remove all 10-K forms that lack necessary stock data (e.g., stocks' daily closing price around the day of filing of 10-K form) from Compustat and CRSP databases. Our final sample is composed of 7,679 firm-year observations of 1,924 unique firms ranging from 2006 to 2010. To avoid an unbalanced sample, we ensure that each firm has at least three year observations. Table 7 presents summary statistics of selected variables in our sample. In particular, the risk type variables (*topic*1 to *topic*30) are discovered and quantified using our sent-LDA-VEM model. The description of each risk type is shown in Figure 6; *SRVA* (stock return volatility after filing) is the measure of postdisclosure risk perception, and *SRVB* (stock return volatility before filing) is the measure of predisclosure risk perception.

Since we intend to use the fixed effect model later, we need to verify whether the number of risk types of individual firms change over time. To this end, we plot the heat map of the count matrix of risk types as shown in Figure 9. In the figure, each column corresponds to a company (excluding those with missing values during the five-year observation period), and every five rows (separated by the dashed lines) correspond to a successive five-year observation of the count of a particular risk type. The meaning of the grey scale is indicated in the right side of the figure. We observe that there are indeed time-series variations of risk types for each individual company since there are observable color changes in each column. If there are no time-series variations, there will be no color changes and what we observe will be areas with a single solid color.

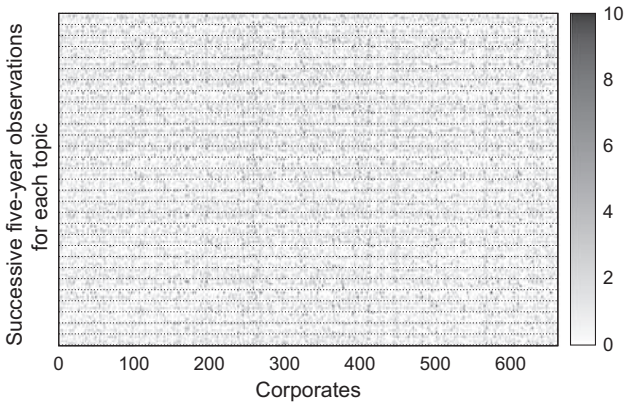
Table 7 Summary Statistics

Variable	Mean	Std. dev.	Min	Max
log *SRVA*	−7.193569	1.309828	−12.91094	2.397478
log *SRVB*	−7.138248	1.236654	−12.0007	1.37464
Price	23.099	30.7174	0.06	767.5
log *Size*	20.14738	1.89597	13.25745	26.14966
log *Trd*	12.34596	2.480582	0	19.97997
Eps	0.7405169	6.935572	493.73	110.36
Wc	137.432	71.29546	0	556
*topic*1	0.8221123	0.7065545	0	5
*topic*2	0.8464644	1.225116	0	9
*topic*3	0.82081	0.9561522	0	13
*topic*4	0.4952468	0.8091989	0	9
*topic*5	0.8669098	1.410339	0	19
*topic*6	0.6322438	0.8943585	0	9
*topic*7	0.6017711	1.371396	0	17
*topic*8	0.9795546	1.216999	0	15
*topic*9	0.430655	1.343795	0	29
*topic*10	0.5349655	0.8099056	0	6
*topic*11	1.008986	0.9760335	0	9
*topic*12	0.8441203	1.191728	0	12
*topic*13	0.5012371	0.9245951	0	11
*topic*14	0.6129704	0.8844219	0	9
*topic*15	0.7318661	1.382401	0	16
*topic*16	0.4777966	1.566448	0	16
*topic*17	0.7330382	2.296263	0	42
*topic*18	1.174632	1.15501	0	14
*topic*19	0.7092069	1.996208	0	23
*topic*20	0.8910014	1.065195	0	11
*topic*21	0.415028	0.8383478	0	13
*topic*22	0.6799062	0.9849882	0	18
*topic*23	1.039849	1.521993	0	14
*topic*24	0.4309155	0.7761337	0	11
*topic*25	1.164344	3.636352	0	33
*topic*26	0.4211486	1.233909	0	30
*topic*27	0.5014976	1.160349	0	12
*topic*28	0.5037114	1.28771	0	20
*topic*29	1.102748	1.144197	0	15
*topic*30	1.016148	2.006242	0	21

Notes. The number of observations is 7,679. *Price*, stock price; log *Size*, market value of equity; log *Trd*, trading volume; *Eps*, earning per share; *Wc*, word counts of disclosures. *Topic*1 to *topic*30 are shown in Figure 6 from left to right, top to bottom.

5.2.2. Econometric Model. We model the influence of different types of risk disclosures on investors' postdisclosure risk perceptions. We estimate a fixed

Figure 9 Heat Map of the Count Matrix of Risk Types



effects linear panel model as shown in the following equation:

$$\begin{aligned}\log SRVA_{it} = {} & \alpha_i + \beta_1 \cdot \log SRVB_{it} + \beta_2 \cdot Price_{it} \\ & + \beta_3 \cdot \log Size_{it} + \beta_4 \cdot \log Trd_{it} + \beta_5 \cdot Eps_{it} \\ & + \beta_6 \cdot Wc_{it} + \beta_T \cdot RiskDisclosure_{it} + \varepsilon_{it}, \quad (2)\end{aligned}$$

where α_i captures unobserved firm specific effects, ε_{it} is the residual random error term, and βs are the model coefficients of interest. In particular, $RiskDisclosure_{it}$ is the set of 30 risk types and β_T is the set of corresponding coefficients. To proxy investors' postdisclosure risk perception, we use the stock return volatility $SRVA_{it}$ of firm i during two months after the filing day in year t.

To obtain unbiased estimates of the effect of risk factor disclosures on investors' risk perception, we control for important relevant factors at different levels, including (1) $SRVB_{it}$, firm i's stock return volatility during the two months prior to the filing of risk disclosures in year t; (2) $Price_{it}$, firm i's closing stock price at the filing day of risk disclosures in year t; (3) $\log Size_{it}$, the log of firm i's market value of equity at the filing day of risk disclosures in year t; (4) $\log Trd_{it}$, the log of firm i's trading volume at the filing day of risk disclosures in year t; (5) Eps_{it}, firm i's earning per share at the filing day of risk disclosures in year t; (6) Wc_{it}, the word counts of textual risk disclosures of firm i in year t; and (7) a set of time dummies at the yearly level and a set of industry dummies based on firms' standard industrial classification code.

5.3. Results

We first estimate a fixed effects (FE) model of investors' postdisclosure risk perceptions (*SRVA*) on all control variables. This baseline model is presented in column (3) of Table 8. As can be seen, the control variables have some explanatory power and their coefficients have the expected signs. Specifically, investors' predisclosure risk perceptions *SRVB*, the log of firms' size ($\log Size$), and the log of firms' trading volume ($\log Trd$) are significantly associated (positively, negatively, and positively, respectively) with postdisclosure risk perception. More importantly, we find that the word counts of risk disclosures are significantly and positively associated with the postdisclosure risk perception. Particularly, an additional unique word is associated with a 0.11% increase in investors' postdisclosure risk perception. This finding is consistent with previous studies by Campbell et al. (2014) and Kravet and Muslu (2011).

We next estimate a full fixed effects model by further including all risk types. This full model is presented in column (1) of Table 8. We choose the FE model rather than a random effects (RE) model because the Hausman test suggests that the RE estimates are inconsistent ($\chi^2 = 221.89$, $p = 0.000$).

To test the joint significance of our discovered risk type variables (*topic*1 to *topic*30), we conduct a likelihood ratio test on the nested models FE full and FE-controls in columns (1) and (3) of Table 8. The result of likelihood ratio test ($\chi^2[30] = 85.36$, $p = 0.000$) shows the joint significance of our risk type variables. More importantly, we have previously demonstrated that our sent-LDA-VEM model could find incremental information (risk types that are ignored by Huang and Li 2011 when predefining the categories for supervised learning methods). To test the joint significance of our additional risk types, we conduct another likelihood ratio test on two nested models—the FE full model and the FE full model excluding 8 risk types (i.e., the new risk types in Figure 6) that are not in 25 risk types defined by Huang and Li (2011). The result of the likelihood ratio test ($\chi^2[8] = 18.95$, $p = 0.015$) demonstrates the joint significance of our incremental risk type variables. This implies that adding our newly discovered risk types as predictor variables results in statistically significant improvement in terms of the model fit.

5.3.1. Discussion on the Effects of Risk Types.

Here, we discuss the effects of risk types on the postdisclosure risk perceptions of investors by interpreting the results of the full FE model as shown in column (1) in Table 8. The mapping of topic ID to risk type labels has been shown in Figure 6. Interestingly, our findings provide support for all three competing arguments about whether and how risk disclosures affect risk perceptions of investors, as discussed as follows.

Support for Null Argument. First, we find that 22 out of 30 risk types have no significant influence on the postdisclosure risk perceptions of investors. This finding lends support to the null argument that risk disclosures are by and large boilerplate. Indeed, there is a long-standing criticism that risk disclosures in financial reports are unlikely to be informative (Schrand and Elliott 1998). To deal with this issue, SEC has repeatedly called for increased focus and specificity

Table 8 Estimation Results

Variable	(1) FE		(2) RE		(3) FE-controls	
*topic*1	−0.0623**	(0.050)	−0.0458***	(0.006)		
*topic*2	−0.0399	(0.185)	0.0069	(0.588)		
*topic*3	0.0108	(0.629)	0.0106	(0.378)		
*topic*4	−0.0826***	(0.006)	−0.0344**	(0.02)		
*topic*5	−0.0208	(0.261)	−0.0145	(0.125)		
*topic*6	−0.0120	(0.627)	−0.0050	(0.698)		
*topic*7	0.0181	(0.295)	0.0119	(0.223)		
*topic*8	0.0295**	(0.039)	0.0226**	(0.014)		
*topic*9	0.0312	(0.113)	0.0208**	(0.028)		
*topic*10	−0.0155	(0.545)	−0.0087	(0.536)		
*topic*11	−0.0255	(0.263)	−0.0265**	(0.020)		
*topic*12	0.0189	(0.342)	0.0096	(0.354)		
*topic*13	0.0075	(0.748)	0.0080	(0.520)		
*topic*14	0.0551**	(0.015)	0.0326**	(0.012)		
*topic*15	−0.0217	(0.241)	−0.0087	(0.336)		
*topic*16	−0.0118	(0.626)	0.0213*	(0.055)		
*topic*17	0.0107	(0.401)	0.0147**	(0.014)		
*topic*18	−0.0364*	(0.061)	−0.0251**	(0.012)		
*topic*19	−0.0134	(0.367)	−0.0241***	(0.002)		
*topic*20	0.0126	(0.563)	−0.0024	(0.825)		
*topic*21	0.0026	(0.911)	−0.0131	(0.345)		
*topic*22	−0.0106	(0.646)	0.0014	(0.910)		
*topic*23	0.0024	(0.897)	−0.0014	(0.877)		
*topic*24	−0.0039	(0.860)	0.0005	(0.971)		
*topic*25	0.0167	(0.203)	0.0035	(0.507)		
*topic*26	−0.0250	(0.275)	−0.0119	(0.226)		
*topic*27	−0.0410*	(0.056)	−0.0147	(0.164)		
*topic*28	0.0578***	(0.002)	0.0203**	(0.044)		
*topic*29	−0.0108	(0.487)	−0.0224**	(0.018)		
*topic*30	−0.0279*	(0.066)	−0.0105	(0.135)		
log *SRVB*	0.4252***	(0.000)	0.5213***	(0.000)	0.4324***	(0.000)
Price	0.0002	(0.824)	0.0010**	(0.011)	0.0002	(0.858)
log *Size*	−0.4415***	(0.000)	−0.2260***	(0.000)	−0.4513***	(0.000)
log *Trd*	0.0479***	(0.000)	0.0799***	(0.000)	0.0499***	(0.000)
Eps	0.0001	(0.968)	−0.0038***	(0.006)	−0.0001	(0.942)
Wc	0.0017**	(0.025)	0.0009**	(0.018)	0.0011***	(0.002)
Intercept	3.9039***	(0.000)	0.0994	(0.747)	4.0595***	(0.000)
Time dummies	Included		Included		Included	
Industry dummies	Included		Included		Included	
Hausman test	$\chi^2 = 211.89,\ p = 0.000$					

Note. *p*-values in parentheses.
*Significant at the 10% level; **significant at the 5% level; ***significant at the 1% level.

in risk factor disclosures, and warned firms to avoid generic risk factor disclosures that could apply to any company (SEC 2005). To examine whether the efforts of SEC have paid off, some recent studies investigate the impact of risk disclosures in 10-K forms and reject the null argument that they lack informativeness (Kothari et al. 2009, Campbell et al. 2014, Kravet and Muslu 2011). One limitation of these studies is that they cannot drill down into analyzing fine-grained risk types because of the lack of methods for measuring qualitative textual information.

Different from these studies, our finding suggests that around two-thirds (22 out of 30) of the different types of risk disclosures are still not informative enough. For example, *topic*11 (competition risks) is one risk type that is frequently reported by firms, but most of its disclosures are quite uninformative and simply say that the firm "operates in a competitive industry." Another example is that *topic*6 and *topic*12 (volatile stock price risks) do not significantly affect the risk perceptions of investors. This is a little surprising since this risk type should be the exact information that the investors need when making decisions. One possible explanation is that investors do not trust the prediction of future stock performance by the firms themselves, but instead make their own assessments based on other indirect but reliably disclosed information. Although superficially not very useful, all of our insignificant associations shed light on what types of risk disclosures lack informativeness. Accordingly, regulators like SEC could make new policy for requiring firms to increase their informativeness.

Support for Divergence Argument. Second, we find that 3 out of 30 risk types, including *topic*8 (macroeconomic

risks), *topic*14 (funding risks), and *topic*28 (credit risks), are positively associated with the postdisclosure risk perceptions of investors. Specifically, an additional sentence of disclosure about "macroeconomic risks," "funding risks," and "credit risks" will lead to a 2.95%, 5.51%, and 5.78% increase of the postdisclosure risk perceptions at 5%, 5%, 1% significant levels, respectively.

The results suggest that the forward-looking statements (in section 1A of 10-K form) about the systematic risks (i.e., macroeconomic risks) are informative, and will increase the postdisclosure risk perceptions of investors (measured by stock return volatility), even if the source of disclosure is the firm itself. This might be due to the prior evidence that systematic (economic-wide) risks cannot be eliminated through diversification, and thus the investors should incorporate this risk into firm value (Fama and French 1993). The results also suggest that the disclosures of liquidity risks (i.e., funding risk and credit risk that may be compounded by liquidity risk) are informative and will increase the postdisclosure risk perceptions of investors. This is consistent with the prior evidence that liquidity-related forward-looking statements have more predictive power in forecasting future liquidity situations and future earnings (Li 2010a).

This finding is partially consistent with recent studies (Campbell et al. 2014, Kravet and Muslu 2011) that support the divergence argument for the risk disclosures in 10-K form. In particular, Campbell et al. (2014) found a positive association between the length of risk disclosures and postdisclosure market-based assessment of firm risk, suggesting that investors incorporate information conveyed by risk disclosures into their assessments of firm risk and stock price. Kravet and Muslu (2011) found that annual increases in risk disclosures are associated with increased stock return volatility around and after the filings, suggesting that risk disclosures increase the risk perceptions of investors. Different from those previous studies, we identify the specific risk types that have impacts rather than mix them together.

Support for Convergence Argument. Third, we find that 5 out of 30 risk types, including *topic*1 (human resources risks), *topic*4 (regulation changes: shareholders' interests), *topic*18 (regulation changes: environment), *topic*27 (infrastructure risks: information security), and *topic*30 (infrastructure risks: disruption), are negatively associated with the postdisclosure risk perceptions of investors. At 1% significance level, an additional sentence of disclosures about "regulation changes: shareholders' interests" will lead to a 8.26% decrease of the postdisclosure risk perceptions, respectively. At 5% significance level, an additional sentence of disclosures about "human resources risks" will lead to a 6.23% decrease of the postdisclosure risk perceptions. At 10% significant level, an additional sentence of disclosures about "regulation changes: environment," "infrastructure risks: information security," and "infrastructure risks: disruption" will lead to a 3.64%, 4.10%, and 2.79% decrease of the postdisclosure risk perceptions, respectively.

Interestingly, these risk types (i.e., human resources, infrastructure, and regulation changes) are unsystematic (i.e., firm-specific or industry-specific) risks. Since the unsystematic risks can be diversified, some prior studies (Campbell et al. 2014) argue that investors should not react as strongly to systematic risks that cannot be diversified. Our results suggest that the informative disclosure of certain unsystematic risks will even decrease the postdisclosure risk perceptions of investors. One possible explanation might be that those legal risks (i.e., regulation changes) and firm-specific risks (i.e., human resources and infrastructure) could reduce the information difference across investors by increasing the quantity of public information. As suggested by Easley and O'Hara (2004), private information increases the risk to uninformed investors of holding the stock, and firms can reduce their cost of capital by affecting the precision and quantity of information available to investors.

This finding is contradictory to recent studies (Kothari et al. 2009, Campbell et al. 2014, Kravet and Muslu 2011) that lend support to the divergence argument for the risk disclosures in 10-K form. The reason is probably that those previous studies cannot drill down into the fine-grained risk types, and thus cannot discover these risk type specific relationships.

5.3.2. Comparison of Competing Methods in Terms of Econometric Model Fit. As shown in the previous section, original LDA models and local-LDA models are less effective than our proposed sent-LDA-VEM model in terms of both statistical fit and substantive fit. Although the topics generated by the less effective methods might be not meaningful for representing risk types, we test whether these topics (as variables in the econometric model) could lead to a better econometric model fit.

To assess the model fit, we choose to use Bayesian information criterion (BIC) statistics rather than pseudo R^2. This is because BIC penalizes for including variables that do not significantly improve fit and allows the comparisons of both the nested and nonnested models (Raftery 1995). In particular, we run the ordinary least squares (OLS) model using the same dependent and independent variables listed in column (1) of Table 8. We run the model three times where, on each occasion, we generate 30 topics by using LDA-CGS, local-LDA-CGS, and our proposed sent-LDA-VEM model, respectively. The most effective learning algorithm is chosen for each model. Table 9 reports the BIC model fit for all three topic models. According to the guidelines stipulated by

Table 9 BIC Differences Between OLS Models Using Risk Type Variables Inferred by Different Models

	BIC	BIC difference with (3)	Evidence
(1) LDA-CGS	−7,254.387	8.036	Strong support for (3)
(2) Local-LDA-CGS	−7,257.063	5.360	Positive support for (3)
(3) Sent-LDA-VEM	−7,262.423	—	—

Note. The guideline by Raftery (1995) for interpreting the magnitude of absolute BIC difference: weak (0–2), positive (2–6), strong (6–10), very strong (>10).

Raftery (1995), the difference of 8.036 in BIC between LDA-CGS and sent-LDA-VEM provides the strong support for our sent-LDA-VEM model; and the difference of 5.360 in BIC between local-LDA-CGS and sent-LDA-VEM provides the positive support for our sent-LDA-VEM model.

6. Concluding Remarks

In this paper, we introduce unsupervised learning methods into the field of financial accounting. To simultaneously discover and quantify risk types from textual disclosures, we propose a novel unsupervised topic model, called sent-LDA, with a variational EM learning algorithm. We conduct comprehensive evaluations in terms of both statistical fit and substantive fit. Experimental results show that our proposed sent-LDA-VEM method outperforms all competing unsupervised methods, and could find more meaningful risk types. By taking advantage of our sent-LDA-VEM method for measuring risk types from textual disclosures, we conduct an empirical study to investigate the effects of risk disclosures on the postdisclosure risk perceptions of investors. Different from prior studies, our findings provide support for three competing arguments regarding whether and how risk disclosures affect the risk perceptions of investors, depending on the specific risk types disclosed. Specifically, we find that around two-thirds of risk types lack informativeness and have no significant influence. Moreover, we find that the informative risk types do not necessarily increase the risk perceptions of investors—the disclosure of three types of systematic and liquidity-related risks will increase the risk perceptions of investors, and the other five types of unsystematic risks will decrease them.

The findings of our empirical study have practical implications for managers and regulators. First, our findings provide managers with more precise understanding on the effects of risk disclosures at the individual risk type level. Although risk disclosures are generally pessimistic, they do not necessarily increase investors' postdisclosure risk perceptions. Besides, since the disclosures by the firm itself can have a significant impact, managers could take active measures to influence investors by carefully choosing the quantity of each risk types to disclose. Second, our findings show that one-third of the disclosed risk types are informative, whereas the other two-thirds lack informativeness. This challenges the findings of prior studies that the disclosures in the newly added section 1A of 10-K forms (regulation S-K, item 503(c)) is informative in general. Since our empirical study sheds some light on what types of risk disclosures lack informativeness, regulators like SEC could make corresponding policies for requiring firms to improve the informativeness of those disclosures.

This study is not without limitations. First, it is still a challenge to evaluate unsupervised clustering methods for exploratory analysis that is used to discover patterns in the data. Von Luxburg et al. (2012) suggest that unsupervised clustering methods should always be studied and evaluated in the context of their end use. We take a small step in this direction by proposing and evaluating an unsupervised topic model for the exploratory analysis of corporate risk disclosures. However, it is still an open question for designing more robust evaluation methods so that the user could trust the model describing text that he or she has never read. Second, we do not explore the correlations between the inferred risk types. Clearly, some risk types are more related with each other and might be merged together. It would be useful to explore these correlations and perform a hierarchical clustering of the learned topics so that we could obtain a more concise taxonomy of risk types. Third, our empirical study only examines the effects of risk disclosures on the postdisclosure risk perceptions of investors. There are many other dependent variables that could be investigated, such as future earnings, information asymmetry, and so on.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/mnsc.2014.1930.

Acknowledgments

The authors thank the department editor, an anonymous associate editor, three anonymous referees, and the participants at the 2012 International Conference on Information Systems for their valuable comments and suggestions.

References

Aral S, Ipeirotis P, Taylor S (2011) Content and context: Identifying the impact of qualitative information on consumer choice. Galletta DF, Liang TP, eds. *32nd Internat. Conf. Inform. Systems* (AIS, Atlanta), 1–9.

Asuncion A, Welling M, Smyth P, Teh YW (2009) On smoothing and inference for topic models. Blimes J, Ng AY, eds. *The 25th Conf. Uncertainty in Artificial Intelligence* (AUAI Press, Corvallis, OR), 27–34.

Azzopardi L, Girolami M, Van Risjbergen K (2003) Investigating the relationship between language model perplexity and IR precision-recall measures. *The 26th Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval* (ACM, New York), 369–370.

Blei DM, Jordan MI (2006) Variational inference for Dirichlet process mixtures. *Bayesian Anal.* 1(1):121–143.

Blei DM, Lafferty JD (2007) A correlated topic model of *Science*. *Ann. Appl. Statist.* 17–35.

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J. Machine Learn. Res.* 3:993–1022.

Brody S, Elhadad N (2010) An unsupervised aspect-sentiment model for online reviews. *The 11th Annual Conf. North Amer. Chapter of the Assoc. Comput. Linguistics* (ACL, Stroudsburg, PA), 804–812.

Campbell JL, Chen HC, Dhaliwal DS, Lu HM, Steele LB (2014) The information content of mandatory risk factor disclosures in corporate filings. *Rev. Accounting Stud.* 19(1):396–455.

Cecchini M, Aytug H, Koehler GJ, Pathak P (2010) Making words work: Using financial text as a predictor of financial events. *Decision Support Systems* 50(1):164–175.

Chang J, Boyd-Graber JL, Gerrish S, Wang C, Blei DM (2009) Reading tea leaves: How humans interpret topic models. Bengio Y, Schuurmans D, Lafferty JD, Williams CKI, Culotta A, eds. *Adv. Neural Inform. Processing Systems* (Curran Associates, New York), 288–296.

Chang YL, Chien JT (2009) Latent Dirichlet learning for document summarization. *The 34th IEEE Internat. Conf. Acoustics, Speech, and Signal Processing* (IEEE, New York), 1689–1692.

Du L, Buntine W, Jin HD (2010) A segmented topic model based on the two-parameter Poisson-Dirichlet process. *Machine Learn.* 81(1):5–19.

Easley D, O'Hara M (2004) Information and the cost of capital. *J. Finance* 59(4):1553–1583.

Fama EF, French KR (1993) Common risk factors in the returns on stocks and bonds. *J. Financial Econom.* 33(1):3–56.

Feldman R, Govindaraj S, Livnat J, Segal B (2010) Management's tone change, post earnings announcement drift and accruals. *Rev. Accounting Stud.* 15(4):915–953.

Garfinkel JA (2009) Measuring investors' opinion divergence. *J. Accounting Res.* 47(5):1317–1348.

Griffiths TL, Steyvers M (2004) Finding scientific topics. *Proc. Natl. Acad. Sci. USA* 101(Suppl 1):5228–5235.

Grimmer J (2010) A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Anal.* 18(1):1–35.

Grimmer J, King G (2011) General purpose computer-assisted clustering and conceptualization. *Proc. Natl. Acad. Sci. USA* 108(7): 2643–2650.

Grimmer J, Stewart BM (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Anal.* 21(3):267–297.

Huang KW, Li ZL (2011) A multilabel text classification algorithm for labeling risk factors in SEC form 10-K. *ACM Trans. Management Inform. Systems* 2(3):1–19.

Humpherys SL, Moffitt KC, Burns MB, Burgoon JK, Felix WF (2011) Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems* 50(3):585–594.

Jo Y, Oh AH (2011) Aspect and sentiment unification model for online review analysis. Davison BD, Suel T, Craswell N, Liu B, eds. *The 4th ACM Internat. Conf. Web Search and Data Mining* (ACM, New York), 815–824.

Kothari SP, Li X, Short JE (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. *The Accounting Rev.* 84(5):1639–1670.

Kravet T, Muslu V (2011) Textual risk disclosures and investors' risk perceptions. *Rev. Accounting Stud.* 18(4):1088–1122.

Li F (2010a) The information content of forward-looking statements in corporate filings: A naive Bayesian machine learning approach. *J. Accounting Res.* 48(5):1049–1102.

Li F (2010b) Textual analysis of corporate disclosures: A survey of the literature. *J. Accounting Literature* 29:143–165.

Lin CH, He YL, Everson R (2011) Sentence subjectivity detection with weakly-supervised learning. *The 5th Internat. Joint Conf. Natural Language Processing* (ACL, Stroudsburg, PA), 1153–1161.

Linsmeier TJ, Thornton DB, Venkatachalam M, Welker M (2002) The effect of mandated market risk disclosures on trading volume sensitivity to interest rate, exchange rate, and commodity price movements. *Accounting Rev.* 77(2):343–377.

Lu B, Ott M, Cardie C, Tsou BK (2011) Multi-aspect sentiment analysis with topic models. Spiliopoulou M, Wang H, Cook DJ, Pei J, Wang W, Zaïane OR, Wu X, eds. *The 11th IEEE Internat. Conf. Data Mining Workshops* (IEEE, New York), 81–88.

Mei QZ, Shen XH, Zhai CX (2007) Automatic labeling of multinomial topic models. Berkhin P, Caruana R, Wu X, eds. *The 13th ACM SIGKDD Internat. Conf. Knowledge Discovery and Data Mining* (ACM, New York), 490–499.

Mirakur Y (2011) Risk disclosure in SEC corporate filings. Working paper, University of Pennsylvania, Philadelphia. Accessed October 1, 2013, http://repository.upenn.edu/wharton_research_scholars/85.

O'Connor B, Bamman D, Smith NA (2011) Computational text analysis for social science: Model assumptions and complexity. *Proc. 2nd Workshop Comput. Soc. Sci.* Accessed March 22, 2014, http://people.cs.umass.edu/wallach/workshops/nips2011css/papers/OConnor.pdf.

Quinn KM, Monroe BL, Colaresi M, Crespin MH, Radev DR (2010) How to analyze political attention with minimal assumptions and costs. *Amer. J. Political Sci.* 54(1):209–228.

Raftery AE (1995) Bayesian model selection in social research. *Sociol. Methodology* 25:111–164.

Rajgopal S (1999) Early evidence on the informativeness of the SEC's market risk disclosures: The case of commodity price risk exposure of oil and gas producers. *Accounting Rev.* 74(3):251–280.

Rogers J, Van Buskirk A, Zechman S (2011) Disclosure tone and shareholder litigation. *The Accounting Rev.* 86(6):2155–2183.

Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20:53–65.

Schrand CM, Elliott JA (1998) Risk and financial reporting: A summary of the discussion at the 1997 AAA/FASB conference. *Accounting Horizons* 12:271–282.

SEC (2005) Securites and Exchange Commission final rule, release no. 33-8591 (FR-75). http://www.sec.gov/rules/final/33-8591.pdf.

Shalen CT (1993) Volume, volatility, and the dispersion of beliefs. *Rev. Financial Stud.* 6(2):405–434.

Sim J, Wright C (2005) The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy* 85(3):257–268.

Tausczik YR, Pennebaker JW (2010) The psychological meaning of words: LIWC and computerized text analysis methods. *J. Language Soc. Psych.* 29(1):24–54.

Teh YW, Kurihara K, Welling M (2008) Collapsed variational inference for HDP. *Adv. Neural Inform. Processing Systems* 20(20):1481–1488.

Teh YW, Newman D, Welling M (2007) A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. *Adv. Neural Inform. Processing Systems* 19:1353–1360.

Titov I, McDonald R (2008) Modeling online reviews with multi-grain topic models. *The 17th ACM Internat. Conf. World Wide Web* (ACM, New York), 111–120.

Von Luxburg U, Williamson RC, Guyon I (2012) Clustering: Science or art? *J. Machine Learn. Res.* 27:65–80.

Wallach HM, Mimno D, McCallum A (2009) Rethinking lda: Why priors matter. *Adv. Neural Inform. Processing Systems* 22:1973–1981.

Wallach HM, Jensen ST, Dicker LH, Heller KA (2010) An alternative prior process for nonparametric Bayesian clustering. *J. Machine Learn. Res.* 9:892–899.

Wang DD, Zhu SH, Li T, Gong YH (2009) Multi-document summarization using sentence-based topic models. *The 4th Internat. Joint Conf. Natural Language Processing* (ACL, Stroudsburg, PA), 297–300.

Zhai K, Boyd-Graber J, Asadi N, Alkhouja ML (2012) Mr. LDA: A flexible large scale topic modeling package using variational inference in MapReduce. Mille A, Gandon FL, Misselis J, Rabinovich M, Staab S, eds. *The 21st ACM Internat. Conf. World Wide Web* (ACM, New York), 879–888.