SUMMARIZATION OF CORPORATE RISK FACTOR DISCLOSURE THROUGH TOPIC MODELING

Completed Research Paper

Yang Bao
Department of Information Systems, National University of Singapore
15 Computing Drive, #COM2-01-02, Singapore 117418
baoyang@comp.nus.edu.sg

Anindya Datta
Department of Information Systems, National University of Singapore
15 Computing Drive, #COM2-04-14, Singapore 117418
datta@comp.nus.edu.sg

Abstract

In this paper, we propose a novel problem of summarizing textual corporate risk factor disclosure, which aims to simultaneously infer the risk types across corpus and assign each risk factor to its most probable risk type. To solve the problem, we develop a variation of LDA topic model called Sent-LDA. The variational EM learning algorithm, which guarantees fast convergence, is derived and implemented for our model. Experiments show that our model is much more efficient and effective than LDA for solving our proposed problem. Specifically, our model is 50 times faster than LDA in the same conditions, and generates better topics for summarization than LDA. Our model is visualized in a publicly available system.

Keywords: Risk factor disclosure, summarization, topic modeling, variational EM
Introduction

The annual report of a corporation is an important public information source 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 such reports by corporations 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 produce an annual report in a specific format, encapsulated in the 10-K form, each year. In addition to the quantitative financial data present in these reports, one of the most significant parts in the 10-K form is the textual risk factor disclosures about the corporation, since stakeholders are particularly sensitive to risks. These risk disclosures are considered so important, that starting in 2005, the SEC requires (as do virtually all major regulatory agencies globally) all firms to include a separate section in their 10-K form to discuss "the most significant factors that make the company speculative or risky" (Regulation S-K, Item 305(c), SEC 2005). This section has turned out to be one of the most examined and debated segments of corporate annual reports (Campbell et al., 2010).

In spite of the significant attention paid to these disclosures, they have widely publicized readability issues (Loughran and McDonald). Several factors contribute to this. First, these textual risk factor disclosures are typically quite lengthy and full of legalese, rendering them hard to decipher for, say, the typical individual investor. Take the large media research firm Nielsen's 10K submission in 2011 1 for example, which details 28 distinct risk factors in section 1A. Each of them is a summary heading, followed by additional explanations. Below we include just a small segment from one of them to illustrate the obtuseness referred to above:

"Continued adverse market conditions, particularly in the consumer packaged goods, media, entertainment, telecommunications or technology industries in particular, could adversely impact our revenue." "A number of adverse financial developments continue to impact the U.S. and global financial markets. These developments include a significant economic deterioration both in the United States and globally, volatility and deterioration in the equity markets, and deterioration and tightening of liquidity in the credit markets. In addition, issues related to sovereign debt in Europe recently have negatively affected the global financial markets. The current economic environment has witnessed a significant reduction in consumer confidence and demand, impacting the demand for our customers' products and services. Those reductions could adversely affect the ability of some of our customers to meet their current obligations to us and hinder their ability to incur new obligations until the economy and their businesses strengthen. The inability of our customers to pay us for our services and/or decisions by current or future customers to forego or defer purchases may adversely impact our business, financial condition, results of operations, profitability and cash flows and may continue to present risks for an extended period of time. We cannot predict the impact of economic slowdowns on our future financial performance ......"

Second, firms describe their risk factors in free-form text, making it quite difficult to categorize, and summarize, the risk factors at the corpus (a collection of all 10-K forms) level. This is particularly vexing for journalists, researchers and interested parties (e.g., insurance companies) who are interested in risks of a particular type. For instance, researchers often search for companies that disclose a specific type of risk factor, say climate changes (Doran and Quinn, 2008), or natural disaster, residential real-estate prices, or federal interest rate fluctuations. To do this, he/she typically has to read through the entire collection of 10-K forms and manually interpret obtuse text. Third, such categorizations are made particularly difficult by virtue of the absence of taxonomy, due to the completely free form nature of these textual disclosures.

Driven by these needs, there has been significant interest, lately, on categorizing risk factors. The bulk of this work has been manual, performed mostly by researchers in social science disciplines, particularly by management researchers. An excellent example of this work is Mirakur et al. (2011) who manually categorized 29 risk types for 122 randomly selected firms. This sample is quite small, compared with the total number of firms that publish 10K forms, but the work quality is thorough. Unfortunately, such

manual efforts are inadequate to work at the corpus level owing to the resource (personnel, time) consuming nature of this work. This naturally leads to the requirement of creating automated methods to solve this problem. This work is at its infancy. The only relevant work we know of, is reported in (Huang and Li, 2011). They proposed a multi-label text classification algorithm to automatically categorize risk factors reported in section 1A of 10-K forms into 25 risk types. However, they assume that all risk types are known before categorization. In other words, they have to manually predefine the risk types by reading hundreds of annual reports. Secondly, their algorithm is supervised and thus need substantial apriori efforts to label the training data.

In this paper, we report the first work on automatic discovery of risk categories, or risk types, and the automatic mapping of each risk factor to a risk type. Risk types may be regarded as a higher level summarization (or categorization) of risk factors. To get a feel for our solution, consider Apple Inc.’s 10-K form in 2006. From section 1A of this form, the summary headings of the first 3 risk factors are listed in Table 1 (for now we ignore the detailed explanation that follows the headings). In this work, 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: Regulation (RT1), Lawsuits (RT2) and Financial Condition (RT3). Further it would map the first risk factor in Table 1 to RT1, the send factor to RT2 and the last factor to RT3. In other words, in this paper, we aim to: (1) automatically discover the risk types in a collection of risk factors in 10-K forms, and (2) automatically assign each risk factor to its most probable risk type.

<table>
<thead>
<tr>
<th>Table 1. The First 3 Risk Factors Disclosed in Section 1A of Apple Inc.’s 10K Form in 2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
</tr>
<tr>
<td>Unfavorable results of legal proceedings could adversely affect the Company’s results of operations.</td>
</tr>
<tr>
<td>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.</td>
</tr>
</tbody>
</table>

In contrast with supervised methods which require the human defined taxonomy of risk factors, our model makes the realistic assumption that predefined risk types are unavailable. This assumption will greatly facilitate the application of our method, making it easily be applied in other similar problems without additional manual work. To discover risk types in an unsupervised fashion, it seems natural to employ unsupervised topic models (Blei et al., 2003) – these methods are capable of inferring topics latent in a collection of documents. Our idea is that by feeding textual risk factors to these methods, it would be possible to discover the underlying risk types in a corpus assuming the inferred topics are representative of risk types. However, it turns out that the standard LDA (Latent Dirichlet Allocation) method (Blei et al., 2003) cannot yield satisfied topics that are semantically representative for risk types. In particular, our experiments show that LDA model will infer many semantically overlapped topics and the meaning of some topics are difficult to be interoperated based on the top probable words. This is due to the fact that LDA enforces a strict bag-of-words assumption which completely ignores the word order in a document. Based on our observation, words in a sentence are usually semantically related and regarding the same topic. We thus propose an extension called Sent-LDA which relaxes the bag-of-words assumption by imposing a constraint that all words in a sentence are generated from one topic.

The main contributions of this paper are summarized as follows.

- We propose a novel problem of summarizing textual corporate risk factor disclosure, which aims to simultaneously infer the risk types across corpus and assign each risk factor to its most probable risk type.

---

2 http://www.sec.gov/Archives/edgar/data/320193/0001104659-06-084288.txt

3 For convenience, we use the terminology “topic” and “risk type” interchangeably in the rest part of the paper.
We propose a variation of LDA topic model called Sent-LDA, which could solve our problem efficiently and effectively. We derive and implement its variational EM learning algorithm which guarantees the fast convergence.

Experiments show that: (1) our Sent-LDA model is 50 times faster than LDA model in the same conditions; and (2) our Sent-LDA model could generate more semantically meaningful topics (risk types) for our summarization task than LDA.

Instead of merely concentrating on testing the performance of proposed solutions in many similar works, we further visualize our learned model in a publicly available system 4. We hope our system could not only help readers browse the textual risk factor disclosure more easier, but also provide a tool for researchers to conduct the textual analysis of risk factor disclosure more effectively.

There are many practical applications of our proposed method. For instance, financial analysts and business managers could easily acquire the risk disclosure information about any interested risk types or any interested corporations without the need to read large amount of textual reports. Researchers could identify and quantify the interested variables (e.g. risk types) from large amount of unstructured textual data, which is seldom utilized in social science researches (e.g. econometrics) due to the lack of computational techniques.

Although our method is utilized for textual risk factor disclosure summarization in this paper, it could be easily applied in other similar problems (e.g. opinion summarization from user generated textual online reviews) without any additional manual work. Compared with supervised solutions which require the manual labeling, the unsupervised nature of our method greatly facilitates its applications such as the exploration of textual data for social science researchers.

The rest of the paper is organized as follows. In the next section, we review some related works. In the follow-up section, we describe our data and methods. We then present two applications of our method. The final section concludes the paper and presents some future works.

Related Works

Our work belongs to the well known research area of computational text analysis (CTA), which aims to quantify the semantic content of textual information. As surveyed by (Connor et al., 2011), there is an increasing interest in the use of CTA in the services of social science questions. They argue that automatic content analysis of texts, 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, computational text analysis has been applied in many fields of social science, including political science, economics, psychology, history and so on (Connor et al., 2011).

While this work is the first to apply CTA to risk factor disclosures, there is much prior work on applying it on other areas of corporate reporting. Li (2011) provides an excellent survey for this line of research. In this survey, the author categorizes this literature based on the economic questions they examine, including: (1) the implications of corporate textual disclosures for earnings quality; (2) valuation of textual information and test of market efficiency; (3) textual disclosure and firms' information environment; (4) textual disclosure and litigation; and (5) implications of corporate textual disclosures for organizational design and corporate financial policies.

Our work is distinct from the above body of work in two major ways: (1) Our work explores an entirely new domain, that of automatic analysis of corporate risk disclosures, and (2) more importantly, we propose a new algorithmic methodology of content analysis of textual corporate disclosures using computational algorithms rather than the (often statistical) exploration of relationships between the content of textual disclosures and various dependent variables. In Table 2, we summarize the computational textual analysis papers according to their methodologies by extending appendix A2 in (Li, 2010) and appendix A in (Campbell et al., 2010).

4 http://www.comp.nus.edu.sg/~baoyang/10kslda/browse/topic-list.html
## Table 2. Summarization of Textual Analysis Papers

<table>
<thead>
<tr>
<th>Paper</th>
<th>Text Analyzed</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Davis et al. (2011)</td>
<td>Press releases</td>
<td>Classify words as optimistic/pessimistic</td>
</tr>
<tr>
<td>Tetlock et al. (2007)</td>
<td>News articles</td>
<td>Classify words as optimistic/pessimistic</td>
</tr>
<tr>
<td>Nelson and Pritchard (2007)</td>
<td>MD&amp;A</td>
<td>Classify words based on type of risk</td>
</tr>
<tr>
<td>Henry (2008)</td>
<td>Earning release</td>
<td>Classify words as positive/negative</td>
</tr>
<tr>
<td>Li (2008)</td>
<td>10Ks and different sections</td>
<td>Classify at word level and averaged across sentences</td>
</tr>
<tr>
<td>Matsumoto et al. (2011)</td>
<td>Conference calls</td>
<td>Classify words into forward-looking</td>
</tr>
<tr>
<td>Feldman et al. (2010)</td>
<td>MD&amp;A</td>
<td>Classify words as positive/negative</td>
</tr>
<tr>
<td>Kothari et al. (2009)</td>
<td>MD&amp;A, analyst report, etc.</td>
<td>Classify words as optimistic/pessimistic</td>
</tr>
<tr>
<td>Rogers et al. (2011)</td>
<td>Earnings announcements</td>
<td>Classify words as optimistic/pessimistic</td>
</tr>
<tr>
<td>Henry and Leone (2009)</td>
<td>Earnings releases</td>
<td>Classify words as positive/ negative</td>
</tr>
<tr>
<td>Li (2010)</td>
<td>MD&amp;A</td>
<td>Classify sentences as positive/negative/neutral/uncertain</td>
</tr>
<tr>
<td>Muslu et al. (2010)</td>
<td>MD&amp;A</td>
<td>Classify words into forward-looking</td>
</tr>
<tr>
<td>Cecchini et al. (2010)</td>
<td>MD&amp;A</td>
<td>Classify risk events using ontology</td>
</tr>
<tr>
<td>Humpherys et al. (2011)</td>
<td>MD&amp;A</td>
<td>Classify fraud risks using linguistic features</td>
</tr>
<tr>
<td>Campbell et al. (2010)</td>
<td>Item 1A of Form 10-K</td>
<td>Classify words based on type of risk</td>
</tr>
<tr>
<td>Huang and Li (2011)</td>
<td>Item of Form 10-K</td>
<td>Classify risk factors using supervised classification</td>
</tr>
<tr>
<td>Aral et al. (2011)</td>
<td>Stock recommendations</td>
<td>Characterize topical content using topic modeling</td>
</tr>
<tr>
<td>This paper</td>
<td>Item 1A of Form 10-K</td>
<td>Summarize risk factors using topic modeling</td>
</tr>
</tbody>
</table>

Among these papers, there are two general approaches for computational text analysis: rule-based ("dictionary") approach and machine learning approach. The dictionary approach uses a “mapping” algorithm to read the text and classify the words or phrases in the text into different categories based on some predefined rules (i.e., dictionary). Although the dictionary approach is easy to use for social researchers, it might not be suitable for doing content analysis for corporate disclosures (Li, 2011). The machine learning approach, on the other hand, can take advantage of data to capture characteristics of interest of their unknown underlying probability distribution. For example, Li (2010) uses a Naive Bayesian classifier to classify the tone and content of forward-looking statements in corporate 10K and 10Q filings. Huang and Li (2011) propose a multi-label text classification algorithm to classify risk factors in section 1A of 10K form into 25 risk types. However, they are all supervised classification algorithms which need human efforts to identify the class and label the training data. (Aral et al., 2011) is the only
work which makes use of unsupervised topic models to characterize the topical content of stork recommendation articles. However, they directly use the standard LDA model without any modification and the bag-of-word assumption of LDA might lead to the discovery of non-meaningful topics in some cases.

**Data and Methods**

In this section, we will first briefly describe the collection of our dataset. Next, we present our methods for solving the proposed problem. Specifically, we start by introducing the standard LDA model from which our model extends, and then elaborate our proposed Sent-LDA model and its learning algorithm.

**Data**

10K forms can be downloaded from EDGAR databases on the SEC’s website. Since it will require enormous amounts of storage if we download all 10K forms into disk, we choose to extract and save only the text in section 1A (i.e., the risk disclosure section) from 10K forms online.

It is quite challenging to extract textual risk factors in section 1A from 10K forms because they are highly unstructured -- some are stored in HTML format while the others are stored in TXT format. To make matters worse, each corporation can freely design the layout of the form. For instance, one could highlight text by using bold font, italic font, underline font or capitalized letters. To deal with these issues, we parse the HTML pages into a tree structure and then scrape needed information using some 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 texts in section 1A. Since our heuristic rules depend on the structure of the form text, we might have some mis-extracted contents. But the quality of extraction is acceptable by manually checking our crawled data. Through this process, we obtained our dataset consisting of 14799 documents of risk factor disclosures in section 1A of 10-K forms.

**Our Approach**

We now proceed to outline our approach, which is an extension of the LDA topic model (Blei et al., 2003). We first provide a brief overview of the LDA model and then proceed to detail our proposed method.

**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 Latent Dirichlet Allocation (LDA), which is developed 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 manifest in the probabilistic generative process associated with LDA, the imaginary random process that is assumed to have produced the observed data. Let $M, N, K, V$ be the number of document in a corpus, number of word in a document, number of topics and number of vocabulary size respectively. $Dirichlet(\cdot)$ is a Dirichlet distribution with parameter $\cdot$ and $Multinomial(\cdot)$ is a multinomial distribution with parameter $\cdot$. $\beta_k$ and $\theta_d$ is $V$-dimensional and $K$-dimensional vector respectively. The generative process of LDA, as shown in a directed graphical model in Figure 1, is:

1. For each topic $k \in \{1, ..., K\}$:
   1. Draw a distribution over vocabulary words $\beta_k \sim Dirichlet(\eta)$
   2. For each document $d$:
      1. Draw a vector of topic proportions $\theta_d \sim Dirichlet(\alpha)$

---

5 http://www.sec.gov/edgar/searchedgar/ftpusers.htm
2. For each word $w_{d,n}$ in document $d$,
   1. Draw a topic assignment $z_{d,n} \sim \text{Multinomial}(\theta_d)$
   2. Draw a word $w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$

Since LDA model infer topic labels in the word level rather than sentence (risk factor) level, we have to further calculate the most probable topic for a sentence using the inferred topics for words in the sentence.

**Sent-LDA**

In this subsection, we start by providing the intuition behind the use of topic modeling for our risk type discovery and mapping problem. We then elaborate our proposed Sent-LDA variant and its learning algorithm.

Since we assume the unavailability of predefined risk types in our proposed problem, it is natural to use unsupervised topic models to discover the underlying risk types in corpus if the inferred topics are representative for risk types. However, we find that LDA cannot discover such representative topics due to its bag-of-words assumption which completely ignores the word order in a document. As discussed in some previous works (Wallach, 2006), the bag-of-words assumption may not always be appropriate. In our case, for example, each risk factor is actually a sentence in which all words are usually discussing one risk type. Based on this observation, we propose our Sent-LDA which relaxes the bag-of-words assumption by imposing a constraint that all words in a sentence are generated from one topic. The same model is previously proposed in (Jo and Oh, 2011) for aspect sentiment analysis. The difference is that they use Gibbs sampling for model learning while we use the variational Bayesian inference. We will illustrate this in detail in the next subsection.

Using the same notations for LDA and let $S$ be the number of sentence in a document, the generative process of our Sent-LDA is changed to:

1. For each topic $k \in \{1, ..., K\}$:
   1. Draw a distribution over vocabulary words $\beta_k \sim \text{Dirichlet}(\eta)$
2. For each document $d$:
   1. Draw a vector of topic proportions $\theta_d \sim \text{Dirichlet}(\alpha)$
   2. For each sentence $s$ in document $d$:
      1. Draw a topic assignment $z_{d,s} \sim \text{Multinomial}(\theta_d)$
      2. For each word $w_{d,s,n}$ in sentence $s$:
         1. Draw a word $w_{d,s,n} \sim \text{Multinomial}(\beta_{z_{d,s}})$

---

6 Actually, a risk factor could be consisted of more than one sentence as shown in Table 1. However, we treat these sentences as one sentence in our Sent-LDA model. In other words, “Sent-LDA” actually means “RiskFactor-LDA.”
This is illustrated as a directed graphical model in Figure 2. Note that our Sent-LDA model adds a sentence layer in the original hierarchy of LDA.

Selection of Learning Algorithm

The key inferential problem we need to solve in order to use LDA (and our Sent-LDA) is that of computing the posterior distribution of the hidden variables given the model and the observed document:

\[ p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w)} \]  

Unfortunately, this distribution is intractable to compute in general (Blei et al., 2003). Topic modeling algorithms form an approximation of Equation (1) by forming an alternative distribution over the latent topic structure that is adapted to be close to the true posterior. Topic modeling algorithms generally fall into two categories (Blei, 2011): sampling-based algorithms and variational algorithms. Sampling based algorithms attempt to collect samples from the posterior to approximate it with an empirical distribution. The most commonly used sampling algorithm for topic modeling is collapsed Gibbs sampling proposed in (Griffiths and Steyvers, 2004). Variational methods are a deterministic alternative to sampling-based algorithms. Rather than approximating the posterior with samples, variational methods posit a parameterized family of distributions over the hidden structure and then find the member of that family that is closest to the posterior. Thus, the inference problem is transformed to an optimization problem. One commonly used variational method is variational Bayesian algorithm proposed in (Blei et al., 2003). Although collapsed Gibbs sampling method is easier to implement, (Asuncion et al., 2008) pointed that its stochastic nature causes it to converge more slowly than the deterministic algorithms such as variational Bayesian algorithm. Thus, we choose to use variational Bayesian algorithm rather than collapsed Gibbs sampling algorithm as in (Jo and Oh, 2011) for the learning of our Sent-LDA model.

Approximate Inference

In posterior inference, we compute the conditional distribution of the latent variables given the observed document. This conditional distribution for our Sent-LDA is as same as that of LDA shown in Equation (1). However, the interpretation of vector \( z \) is changed. Specifically, since we only draw topic assignment for each sentence (the words in a sentence share the same topic assignment) rather than each word, \( z \) is now the topic assignments for each sentence \( s \) rather than the topic assignments for each word \( n \).

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 (KL) divergence to the true posterior. In Sent-LDA model, the latent variables are the per-document topic proportion \( \theta \) and the per-sentence topic assignment \( z \). Similar to variational method for LDA, we use the following variational distribution:

\[ q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \prod_{s=1}^{S} q(z_s|\phi_s) \]

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

We now describe how to set the variational parameter \( \gamma \) and \( \phi \) via an optimization procedure. Following Jordan et al. (1999), we bound the log likelihood of a document using Jensen's inequality. By omitting the variational parameter \( \gamma \) and \( \phi \), we have:

\[
\log p(w|\alpha, \beta) = \log \int_{z} p(\theta, z, w|\alpha, \beta) d\theta \\
= \log \int \sum_z p(\theta, z, w|\alpha, \beta) q(\theta, z)/q(\theta, z) d\theta \\
\geq \int \sum_z q(\theta, z) \log p(\theta, z, w|\alpha, \beta)/q(\theta, z) d\theta - \int \sum_z q(\theta, z) \log q(\theta, z) d\theta \\
= E_q[\log p(\theta, z, w|\alpha, \beta)] - E_q[\log q(\theta, z)] \\
= \operatorname{L}(\gamma, \phi; \alpha, \beta)
\]
By expanding the lower bound $L$ using the factorizations of $p$ and $q$, we have:

\[
L(y, \phi; \alpha, \beta) \\
= E_q [\log p(\theta|\alpha)] + E_q [\log p(z|\theta)] + E_q [\log p(w|z, \beta)] - E_q [\log q(\theta)] - E_q [\log q(z)] \\
= \log \Gamma \left( \sum_{j=1}^{K} \alpha_j \right) - \sum_{i=1}^{K} \log \Gamma(\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1) \left( \psi(y_i) - \psi \left( \sum_{j=1}^{K} y_j \right) \right) \\
+ \sum_{s=1}^{S} \sum_{i=1}^{K} \phi_{si} \left( \psi(y_i) - \psi \left( \sum_{j=1}^{K} y_j \right) \right) \\
+ \sum_{s=1}^{S} \sum_{i=1}^{K} \phi_{si} \sum_{n=1}^{N_s} w_{in} \log \beta_{ij} \\
- \log \Gamma \left( \sum_{j=1}^{K} y_j \right) + \sum_{i=1}^{K} \log \Gamma(\gamma_i) - \sum_{i=1}^{K} (\gamma_i - 1) \left( \psi(y_i) - \psi \left( \sum_{j=1}^{K} y_j \right) \right) \\
- \sum_{s=1}^{S} \sum_{i=1}^{K} \phi_{si} \log \phi_{si}
\]

where $\psi$ is the first derivative of the log $\Gamma$ function. 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 LDA in (Blei et al., 2003) lie in the second, third and fifth term due to the additional sentence layer.

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

\[
\phi_{si} \propto \left( \prod_{n=1}^{N_s} \beta_{wn} \right) \exp \left( \psi(y_i) - \psi \left( \sum_{j=1}^{K} y_j \right) \right) \\
\gamma_i = \alpha_i + \sum_{s=1}^{S} \phi_{si}
\]

**Parameter Estimation**

Given a corpus of documents, we aim to find parameters $\alpha$ and $\beta$ that maximize the log likelihood of the observed data. To achieve this purpose, we use a variational EM procedure (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 previous inference subsection. In the M-step, we find the maximum likelihood estimates of parameters $\alpha$ and $\beta$ 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 maximize the lower bound of likelihood with respect to the model parameter, we obtain the M-step update for the multinominal parameter $\beta$:

\[
\beta_{ij} \propto \sum_{d=1}^{M} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \sum_{i=1}^{K} \phi_{dsni} w_{dni}
\]

For the Dirichlet parameter $\alpha$, 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.

**Experiments**

Recall that the two goals of our proposed problem of summarizing textual risk factor disclosures are: (1) infer the risk types across corpus, and (2) assign each risk factor to its most probable risk type. In our
proposed Sent-LDA model, the learned corpus-specific topics correspond to risk types and the document-specific topic assignments for sentences correspond to the risk type assignments for risk factors. In this section, we conduct several experiments to address the following questions.

- How to determine the appropriate number of topics?
- Is our Sent-LDA model more efficient than LDA model?
- How is the quality of the inferred risk types? How is the quality of the risk type assignments for risk factors? Is our proposed Sent-LDA model better than LDA model for our task?

**Determination of Number of Topics**

Our Sent-LDA model is not nonparametric, and we have to set appropriate parameter $K$, i.e., number of topics, for the model. In this paper, we choose the value of $K$ that will result in a model with good generalization performance, and, more importantly, lead to the discovery of explanatory topics for risk types.

We use perplexity as the metric of generalization performance. Perplexity, originally used in language modeling (Azzopardi et al., 2003), is a common criterion of clustering quality that does not require a priori categorizations. The perplexity is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood given the model. A lower perplexity score indicates better generalization of unseen data. Formally, for a test set of $M$ documents, the perplexity is:

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

where $\log p(w_d)$ is approximated with the lower bound $L(y, \phi; \alpha, \beta)$. As shown in Figure 3, the perplexity for 10% held-out test data decreases and tends to converge with the increase of number of topics.

On the other hand, topic modeling is used as an explanatory tool in our task. As pointed by Chang et al. (2009), metrics of model fit such as perplexity are useful for evaluating the predictive model, but do not address the more explanatory goals of topic modeling. Thus, we also need to make the value of $K$ close to the “real” number of risk types discussed in risk factor disclosures. Authors in (Huang and Li, 2011) identify 25 types of risk by reading through large number of risk factor disclosures in 10-K forms. And we use the number 25 as an approximate of the “real” number risk types in our corpus. Finally, we set the value of $K$ in our model to 30 because: (1) the changing rate of perplexity at this point tends to be low, and (2) it is close to our approximate of “real” number of risk types.

**Complexity Analysis**

To compare the efficiency of our Sent-LDA model and LDA model, we firstly analyze the computational complexity of the variational EM algorithm for LDA. The time complexity of E-step is $O(MN^2K)$ 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 a little smaller than $N$. On the other hand, the time complexity of M-step is $O(VK)$ 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 variational EM algorithm for our Sent-LDA model. Our
time complexity is as same as that of LDA except that the time complexity of E-step for our model is \(O(MS^2K)\) where \(S\) is the number of sentences for 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 model is more efficient than LDA model since \(S\) will be definitely much smaller than \(N\). In particular, in the same Linux system with dual 2.80GHz CPU and 4.0GB memory, to train a model with 30 topics against our dataset using the same convergence criterion, it takes 20 minutes for our Sent-LDA model but 16 hours for LDA model.\(^7\)

**Quality of Inferred Topics**

Since our topics (risk types) are inferred automatically based on the word concurrence in textual sentences through topic modeling while Huang and Li (2011)'s risk types are subjectively defined by humans, it is important to check whether our automatically identified risk types make sense (i.e. be representative for risk types). We carefully compare our identified topics (risk types) with that in Huang's work. Specifically, we manually map our 30 topics with 25 risk types defined in their work based on the most probabilistic words in each topic. The comparison is shown in Table 3, where ‘**’ indicates the risk types that are found by us but ignored in Huang’s work, ‘*’ indicates the risk types that are segmented further by us than Huang’s, and ‘+’ indicates the risk types identified by Huang but merged with some other risk types by us. As can be seen, we could find all risk types that are defined in Huang’s work, although some of them are merged into one risk type. For instance, two risk types -- “Disruption of operations” and “Infrastructure risks” -- are merged into one risk type (topic 18). Besides, we identify some more fine-grained risk types than Huang’s work. For example, the risk type “Regulation changes” in Huang’s work is further segmented into “Regulation changes (accounting)” and “Regulation changes (environment)” in our work. More importantly, we find some risk types that are ignored in Huang’s work, including “Stakeholder’s profit”, “Labor cost”, “Management”, “Investment” and so on. These new risk types justify the effectiveness of our proposed Sent-LDA model.

<table>
<thead>
<tr>
<th>Our risk types represented by the most probabilistic words</th>
<th>Huang’s risk types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 0: investment, property, distribution, interest, agreement</td>
<td>Shareholder’s interest risk</td>
</tr>
<tr>
<td>Topic 1: regulation, change, law, financial, operation, tax, accounting</td>
<td>*Regulation changes(accounting)</td>
</tr>
<tr>
<td>Topic 2: gas, price, oil, natural, operation, production</td>
<td>Input prices risks</td>
</tr>
<tr>
<td>Topic 3: stock, price, share, market, future, dividend, security, stakeholder</td>
<td>**Stakeholder’s profit</td>
</tr>
<tr>
<td>Topic 4: cost, regulation, environmental, law, operation, liability</td>
<td>*Regulation changes(environment)</td>
</tr>
<tr>
<td>Topic 5: control, financial, internal, loss, reporting, history</td>
<td>Financial condition risks</td>
</tr>
<tr>
<td>Topic 6: financial, litigation, operation, condition, action, legal, liability, regulatory, claim, lawsuit</td>
<td>*Potential/Ongoing Lawsuits</td>
</tr>
<tr>
<td>Topic 7: competitive, industry, competition, highly, market</td>
<td>Competition risks</td>
</tr>
<tr>
<td>Topic 8: cost, operation, labor, operating, employee, increase, acquisition</td>
<td>**Labor cost</td>
</tr>
<tr>
<td>Topic 9: product, candidate, development, approval, clinical, regulatory</td>
<td>New product introduction risks</td>
</tr>
<tr>
<td>Topic 10: tax, income, asset, net, goodwill, loss, distribution, impairment, intangible</td>
<td>**Accounting, Restructuring risks</td>
</tr>
<tr>
<td>Topic 11: interest, director, officer, trust, combination, share, conflict</td>
<td>**Management</td>
</tr>
</tbody>
</table>

\(^7\) We use David Blei’s implementation of variational EM for LDA: http://www.cs.princeton.edu/~blei/lda-c/index.html
Since the mapping of risk types in Table 3 are solely based on the most probabilistic words in each topic, we further qualitatively check whether the risk factors are correctly summarized (categorized) into corresponding risk type. In other words, we would like to investigate the quality of topics (risk types) by checking whether risk factors are semantically coherent with the assigned topics (risk types). In particular, we conduct a case analysis of Apple Inc. aforementioned and present the sample summarization for Apple Inc.’s risk factor disclosures in 2006 in Table 4. We manually label each topic with the Huang’s risk types as same as Table 3. It should be noted that this labeling is not a must since top words in each topic are quite self-explained. Actually we only use the top 3 words to label each topic in our demo system since they are sufficient for understanding the topic. The labeling in this experiment only aims for the comparison with Huang and Li (2011)’s work. As can be seen, the risk type assignments for risk factors are reasonably accurate. Particularly, the risk factors assigned to new risk types “Insurance” and “Accounting” is semantically coherent. This demonstrates the rationality of these two new risk types and shows that our Sent-LDA model could find risk types that might be otherwise ignored by human.

---

8 The complete output of our model, including all learned topics, topic distributions and risk assignments and so on, are available at http://www.comp.nus.edu.sg/~baoyang/10kslda/exp
<table>
<thead>
<tr>
<th>Topic</th>
<th>[Manual label] Most likely words in topic</th>
<th>Risk factors assigned to topic</th>
</tr>
</thead>
</table>
| 6     | [Potential/Ongoing Lawsuits] financial, litigation, operation, condition, action, material, legal, liability, regulatory, claim, proceeding, investigation, future, security, lawsuit, cost, cash, harm, flow, outcome | ✓ 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.  
✓ Unfavorable results of legal proceedings could adversely affect the Company’s results of operations. |
| 29    | [Catastrophes] operation, natural, condition, financial, facility, disaster, event, terrorist, operating, weather, affected, act, attack, terrorism, war, loss, disruption, cost, material, revenue | ✓ 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. |
| 20    | [Suppliers risks] [Downstream risks] customer, product, revenue, sale, supplier, operation, material, loss, relationship, supply, operating, limited, financial, component, key, portion, contract, manufacturing, rely, service | ✓ 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.  
✓ The Company is dependent on manufacturing and logistics services provided by third parties, many of whom are located outside of the U.S.  
✓ The Company’s future operating performance is dependent on the performance of distributors and other resellers of the Company’s products. |
| 21    | [Intellectual property risks] property, intellectual, protect, product, proprietary, technology, patent, claim, party, third, litigation, infringement, infringe, adequately, competitive, costly, protection, harm, license, compete | ✓ 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. |
| 15    | [Volatile stock price risks] stock, price, operating, stockholder, fluctuate, interest, quarterly, volatile, fluctuation, decline, control, market, cause, ha, significantly, future, director, influence | ✓ The Company expects its quarterly revenue and operating results to fluctuate for a variety of reasons.  
✓ The Company’s stock price may be volatile. |
| 14    | [Human resource risks] personnel, key, retain, attract, management, employee, loss, qualified, operation, success, service, dependent, senior, executive, skilled, team, hire, lose, officer | ✓ The Company’s success depends largely on its ability to attract and retain key personnel. |
| 28    | {Insurance} loss, insurance, financial, loan, reserve, operation, estimate, condition, credit, actual, allowance, future, increase, liability, claim, assumption, rate, cover, cost, coverage | 
The Company is subject to risks associated with the availability and coverage of insurance.

**International risks** operation, international, foreign, currency, rate, fluctuation, exchange, economic, political, sale, financial, associated, condition, market, operating, revenue, country

The Company’s business is subject to the risks of international operations.

**Regulation changes: environment** cost, regulation, environmental, law, operation, liability, material, increase, compliance, raw, product, operating, price, safety, comply, financial, change

The Company’s business is subject to the risks associated with environmental regulations.

**Regulation changes: accounting** regulation, change, law, financial, operation, tax, accounting, cost, regulatory, government, compliance, standard, condition, comply, act, operating

Changes in accounting rules could affect the Company’s future operating results.

**Accounting** tax, income, asset, net, goodwill, loss, distribution, cash, rate, value, impairment, future, reduce, investment, federal, intangible, charge, reit, unit, required

Changes in the Company’s tax rates could affect its future results.

---

**Comparison between LDA and Sent-LDA**

To quantitatively evaluate the performance of LDA and our Sent-LDA for our proposed problem, we use the **Silhouette Coefficient** (Rousseeuw, 1987), which is an unsupervised metric for measuring the performance of clustering. We choose this unsupervised metric since we have no ground of truth labels for calculating the supervised metrics. For calculating, we use Euclidean distance and 10000 sentences (risk factors). Since LDA model infer topic labels in the word level rather than sentence (risk factor) level, we calculate the most probable topic for a sentence using the inferred topics for words in the sentence. The score is bounded between -1 for bad clustering and +1 for highly dense clustering. The result is shown in Table 5. As can be seen, our Sent-LDA performs better than LDA.

<table>
<thead>
<tr>
<th></th>
<th>LDA</th>
<th>Sent-LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA -0.0891</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sent-LDA -0.0522</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

On the other hand, our proposed problem is not a pure clustering problem in the sense that our problem aims to identify semantically representative topics for risk types rather than to find dense clusters. In fact, topic modeling is a sort of soft-clustering in the sense that it discovers the probabilistic rather than hard associates between text (risk factors) and latent topics (risk types). Each latent topic (cluster) is semantically represented by the multinomial distributions over vocabulary words. Thus, in addition to evaluate the clustering performance, it is also important to compare the quality of topics learned by LDA and Sent-LDA. The good topics should be semantically representative for meaningful risk types. We thus run the LDA model with 30 topics and compare the learned topics with ours. To make the comparison, we first need a way to match two sets of topics learned by two models. This is because topic IDs (0 to 29 if the number of topics is 30) are arbitrary labels which lead to the lack of identifiability. For example, the topic-zero in Sent-LDA might correspond to topic-20 in LDA model. This matching can be done using minimum bipartite matching that minimizes the average KL divergence between topics in two sets (Newman et al., 2006). Specifically, we use the Kuhn-Munkres algorithm (Lovasz and Plummer, 1986), also known as the Hungarian algorithm, for minimum bipartite matching. The average KL divergence between topics in two sets is defined as:
where $KL(P||Q)$ is the symmetric KL divergences and $\text{perm}(t)$ is the permutation that minimizes $avgKL$. We present some matched topics found by two models in Table 6. The KL divergence measures the similarity between two distributions (topics). The lower KL divergence indicates that the two topics are more similar.

As shown in Table 3, the top words in topics generated by Sent-LDA are meaningful enough for mapping the learned topics with risk types defined in Huang’s work. We now check whether the top words in topics generated by LDA are as meaningful as those of Sent-LDA. As shown in Table 6, it is not surprising that those LDA topics (e.g. lda-T5) matched with low KL-divergence with Sent-LDA topics are almost equally meaningful since they are similar. However, for those with higher KL divergence, the topics discovered by LDA are not representative of risk types. For instance, the topic lda-T28 and sent-lda-T29 are very different from each other. The topic sent-lda-T29 is regarding “Catastrophes” while the topic lda-28 is hard to interoperate its semantic meaning based on top words. This means that LDA model could not find the risk type “Catastrophes” because the topics learned by LDA and Sent-LDA are one-to-one matched. By inspecting the full list of topics discovered by LDA, we find many topics that are semantically overlapped or difficult to interpret. Besides, topics discovered by LDA miss most risk types defined by Huang and Li (2011), including “Human resources risks”, “Catastrophes”, “International risks”, “Intellectual property risks”, “Competition risks” and so on. In sum, our Sent-LDA model could discover more representative topics (risk types) than LDA model.

<table>
<thead>
<tr>
<th>KL-divergence</th>
<th>Topic id</th>
<th>Most likely words in topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>49.634838</td>
<td>[sent-lda] T29</td>
<td>operation,natural,condition,financial,facility,disaster,event,terrorist,operation,weather,affected,act,attack,terrorism,war,loss,disruption,cost,material,revenue</td>
</tr>
<tr>
<td></td>
<td>[lda] T28</td>
<td>cost,power,energy,operation,southern,operating,financial,facility,generation,market,capital,change,environmental,rate,electric,regulation,cash,negatively,price,flow</td>
</tr>
<tr>
<td>43.978836</td>
<td>[sent-lda] T14</td>
<td>personnel,key,retain,attract,management,employee,loss,qualified,operation,success,service,dependent,senior,executive,skilled,team,hire,lose,oficer,failure</td>
</tr>
<tr>
<td></td>
<td>[lda] T0</td>
<td>product,revenue,operating,customer,market,sale,harm,stock,price,future,operation,technology,service,fail,cause,intellectual,property,software,decline,financial</td>
</tr>
<tr>
<td>21.940663</td>
<td>[sent-lda] T1</td>
<td>regulation,change,law,financial,operation,tax,accounting,cost,regulator,y,government,compliance,standard,condition,comply,act,operating,federal,requirement,policy,increase</td>
</tr>
<tr>
<td></td>
<td>[lda] T14</td>
<td>financial,operation,condition,change,market,material,materially,affecte,d,economic,credit,negatively,rate,cash,tax,flow,cost,accounting,failure,position,liquidity</td>
</tr>
<tr>
<td>18.932172</td>
<td>[sent-lda]</td>
<td>system,service,information,failure,product,operation,customer,technol</td>
</tr>
</tbody>
</table>

Table 6. Comparison between Topics of LDA and Sent-LDA

---

9 The full match list is available at: [http://www.comp.nus.edu.sg/~baoyang/10kslda/exp/topic_match.hungarian](http://www.comp.nus.edu.sg/~baoyang/10kslda/exp/topic_match.hungarian)

10 [http://www.comp.nus.edu.sg/~baoyang/10kslda/exp/topics.lda](http://www.comp.nus.edu.sg/~baoyang/10kslda/exp/topics.lda)
Applications

In this section, we present two applications of our Sent-LDA model. Firstly, we present a publicly available system which visualizes our learned model with the purpose of helping readers to understand and navigate the risk factor disclosures in 10-K forms. Secondly, we show a simple but interesting content analysis of risk factor disclosures using our model -- the topic drift over time.

Visualization of Learned Sent-LDA Model

It is impractical to organize by hand large collections of documents which are not coherently organized. Probabilistic topic modeling provides a way to automatically discover the hidden thematic structures in the collection and represent each document as a combination of themes. However, as argued by (Chaney and Blei, 2012), topic models are high-level statistical tools -- a user must scrutinize numerical distributions to understand and explore the raw output of those models. They thus develop an open source implementation, called TMVE (topic model visualizing engine), to summarize a collection of documents using topics, reveal the relationships between content (documents) and summaries (topics), and reveal the relationships across contents (documents).
In this paper, we use TMVE to visualize our learned Sent-LDA model and make it publicly available. Since TMVE can only be directly used for LDA model, we modify some components of TMVE for its application on our Sent-LDA model. We present an example in Figure 4 to demonstrate how to use the visualized model to navigate the document collection. Beginning in the upper left, we see a set of topics, each of which is a theme discovered by our topic modeling algorithm. We use top 3 words to label each topic. We click on a topic about “intellectual property protection”. We choose a document associated with this topic, which is the risk factor disclosures of a firm called “E Digital” in 2008. The page about this document includes its content and the topics that it is about. We explore a related topic about “stock price and shares”. Then we can repeat the above process to explore more documents. This browsing structure - the themes and how the documents are organized according to them - is created by running our proposed Sent-LDA on our dataset and visualizing its output. There are some more features of the visualized system such as similar document list, topic distribution for each term and so on. We leave it for readers to further explore 11.

**Topic Drifts over Time**

Our Sent-LDA can also be utilized for various content analyses of textual risk factor disclosures. Here, we only show one simple but interesting content analysis – the drift of topics (risk type) over time. Specifically, financial analysts might be interested in tracking the changes of the number of risk types over time to see if there are any emerging risks at some specific time. As shown in Figure 5, we plot the stacked histograms of the proportions of 30 topics in each year. It is interesting to see that the number of topic 17, which is talking about “Macroeconomic risks” as shown in Table 3, is doubled in 2009. This might be due to the financial crisis in 2009.

**Conclusion and Future Works**

In this paper, we propose a novel problem of summarizing textual corporate risk factor disclosure, which aims to simultaneously infer the risk types across corpus and assign each risk factor to its most probable risk type. To solve the problem, we develop a variation of LDA topic model called Sent-LDA. The variational EM learning algorithm, which guarantees fast convergence, is derived and implemented for our model. Experiments show that our model is much more efficient and effective than LDA for solving our proposed problem. Specifically, our model is 50 times faster than LDA in the same conditions, and generates better topics (risk types) for summarization than LDA. We further demonstrate two applications of our model: (1) we implement a publicly available system by visualizing our learned model to help readers understand and navigate the textual risk factor disclosures in 10-K forms; and (2) conduct a simple but interesting content analysis which aims to track the topic drift over time.

In future, we plan to extend our work in the following aspects. Firstly, we will apply and evaluate our method in other similar text analysis problems. Secondly, we will refine our data collection program in order to obtain better dataset. Thirdly, we could definitely improve the quality of visualization of our model. Finally, we will use some quantitative methods, such as word intrusion and topic intrusion proposed in (Chang et al., 2009), for measuring semantic meaning in inferred topics. Currently, we only qualitative evaluate the quality of learned topics and its assignments to sentences.

---

11 The detailed introduction of our system will be available in our demo paper in preparation.
References


Information Content of Managers' Presentations and Analysts' Discussion Sessions," The Accounting
http://repository.upenn.edu/wharton_research_scholars/85
Community Research and Development.
Tetlock, P.C. 2007. “Giving Content to Investor Sentiment: The Role of Media in the Stock Market,”
Conference on Machine Learning, New York, NY, USA, pp. 977-984.