EXTENDING TOPIC MODELS FOR TEXT ANALYSIS OF CORPORATE RISK DISCLOSURES

BAO YANG

(Bachelor of Management, Nanjing University, 2007)

(Master of Management, Nanjing University, 2009)

A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF INFORMATION SYSTEMS
NATIONAL UNIVERSITY OF SINGAPORE
2013

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Bao Yang

30 September 2013

Acknowledgments

This thesis would not have been possible without the help and support of a multitude of individuals and institutions.

Foremost, I owe my supervisor, Prof. Anindya Datta, many thanks for his mentorship and support in the past four years. He has taught me, both consciously and unconsciously, how good research is done. Without his guidance, expertise, patience, and understanding, the completion of this thesis would not have been possible.

My great gratitude goes to all the other thesis committee members, Prof. Tat-seng Chua, Prof. Danny Chiang-choon Poo, and the external examiner, for their critical reading and valuable feedback. They have devoted much time and effort to help me improving the quality of the thesis. Special thanks go to Prof. Hock-hai Teo, who has served as my oral panel chair, for his valuable suggestions during the thesis defense.

I would also like to thank many other professors who have taught me and supported me at different stages during my Ph.D study, especially, Prof. Jie Zhang, Prof. Nigel Collier, Prof. Kaushik Dutta, Prof. Qinghua Zhu, Prof. Zhenhui Jiang and Prof. Yunjie Xu.

I am fortunate to have a number of great friends who have supported me both research wise and non-research wise, particularly, Fang Fang, Cenzhe Zhu, Xiqing Sha, Jingda Zhan, Haifeng Xu, Xiaoying Xu, Chunmian Ge, Nargis Pervin, and Sangaralingam Kajanan. Without them, my Ph.D journey would have been a less enjoyable one.

I gratefully acknowledge the funding sources that made my Ph.D study possible. I was funded by the NUS research scholarship for my four-year graduate study, and supported by the school of computing with generous travel grants to attend academic conferences for presenting and discussing the works in the thesis. I also thank the National Institute of Informatics (Japan) for hosting an excited five-month internship.

Lastly and most importantly, I would like to thank all of my family members who have given me a lifetime of love and care, and have always been a source of encouragement for me. I give my special thanks to my wife Hui. Her kindness, patience, encouragement, and care sustain me.

To my grandfather, Enpeng Bao.

Table of Contents

Summary

Annual reports submitted by corporations, regulatorily mandated in virtually every country of the globe, comprise one of the most scrutinized classes of documents in the world of corporate finance. In these reports, companies are required to disclose risks that might impact its business in the risk disclosure section of the annual report, where the various risks types facing the company are described in free form text. Such risks are the subject to much analysis in the investment community, particularly by analysts at institutional investment firms and form a basis of the market movement of the company's equity value.

The goals of analyzing risk disclosure text are twofold: (1) to interpret the "overall risk sentiment" embodied in the disclosure section, and (2) to extract the various individual risk types enumerated in the section. Clearly, both of these objectives can be reduced to the solving of text analysis problems. In the first, i.e., risk sentiment identification, complications arise as a result of different industry segments having differing risk disclosure "patterns". In the second case, the existing methods to extract risk types all rely on dictionary based or supervised learning methods which work out to be extraordinarily manual effort intensive.

In this context we provide new solutions to these problems in this thesis. In particular, this thesis is comprised of three studies. First, we study

the problem of cross-industry risk sentiment analysis, where a sentiment classifier is trained on the annotated training data from one industry (i.e., the source industry) but is meant to be used in another (i.e., the target industry). Cross-industry risk sentiment analysis allows the reuse of the annotated training data in source industries, and thus could reduce the amount of manual effort to annotate the training data in the target industry. However, conventional supervised learning methods usually lead to low performance for this problem due to different word "patterns" across industries. We therefore propose an extended LDA (Latent Dirichlet Allocation) topic model, which could bridge the gaps between the source and target industries in the low-level word feature space by learning a new high-level topic feature space in a supervised way. Evaluations are conducted on nine standard testing datasets and one real-world risk disclosure dataset, and the results demonstrate the effectiveness of our proposed method.

Second, we study the problem of extracting various risk types (e.g., funding risk, infrastructure risk, etc.) from textual risk disclosures. To this end, we propose an extended LDA topic model, which could infer topics (i.e., risk types) covered in a set of risk disclosures, and identify the risk types of specific sentences. Different from existing methods, our method does not assume pre-defined categories (i.e., risk types), and thus could reduce the amount of manual effort substantially. We use our model to examine the risk disclosures in 10-K forms from 2006 to 2010. The results demonstrate that our model outperforms all competing methods, and could find more meaningful topics that are representative for risk types. Third, we continue the analysis of risk types extracted in the second study by examining the market reactions

to them. Specifically, we conduct an empirical study to investigate whether and how risk disclosures will affect the post-disclosure risk perceptions of investors at the individual risk type level. Different from prior studies, our results lend support for all three competing arguments on the effects of risk disclosures, depending on the specific risk types disclosed. Our findings have implications for both managers and regulators.

In summary, the main contribution of this thesis is the development of two extended topic models for identifying risk sentiment and extracting various risk types from textual risk disclosures. The proposed methods could facilitate the analysis of corporate risk disclosures by reducing the amount of human effort substantially. Moreover, the proposed methods enable the empirical study of market reactions to risk disclosures at the individual risk type level. The findings of our empirical study reconcile the conflicting arguments about the effects of risk disclosures on post-disclosure risk perceptions of investors in accounting literature.

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

Corporate disclosure is an important way for management to communicate firm performance and governance to various stakeholders, especially outside investors, and is critical to the functioning of an efficient capital market (Healy and Palepu, 2001). There are various sources from which the disclosure is provided, including the regulated financial reports (e.g., corporate annual reports), voluntary communications (e.g., management forecast, conference calls, and press releases), and the information intermediaries (e.g., financial analysts, industry experts, and financial press). Among these different sources, annual reports submitted by firms comprise one of the most scrutinized classes of documents in the world of corporate finance. Due to the importance of these reports, their filing 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). In these reports, companies are required to disclose risks that might impact its business in the risk disclosure section, where the various risks types facing the company are described in free form text. Such risks are believed to be the basis of the market movement of the company's equity value, and are the subject of much analysis by both researchers in the financial accounting community and practitioners in the investment community, particularly analysts at institutional investment firms.

The fundamental problem of analyzing textual risk disclosures is to glean the useful and actionable information from the large amount of unstructured text. Specifically, the goal of this text analysis problem is to categorize the information contained in text into manageable numeric variables of interest. To this end, many prior studies (Li, 2010b) have largely relied on the manual text analysis approaches. Although these approaches can be more precise, they are quite resource (e.g., personnel, time) consuming, as they require manual exhaustive text perusal. Consequently, the follow-up studies are limited to the small-size samples, resulting in many undesirable issues, such as the difficulty with replication and limited generalizability of the empirical results (Li, 2010b).

To mitigate the cost of manual analysis, there is a growing body of research, especially in the financial accounting domain, which adopts automatic text analysis techniques for analyzing textual disclosures (Refer to Section 2.3 for a review of this line of research). These automatic approaches can reduce the amount of human effort to a large extent, and enable the large-sample text analysis. Despite their advantages over the manual approaches, the existing automatic approaches still require high start-up costs of human effort to use. For example, the two most successful and widely adopted types of automatic methods are the dictionary based methods (Loughran and McDonald, 2011) and supervised learning methods Li (2010a). Both of them assume the pre-defined, mutually exclusive, and exhaustive categories (i.e., variables of interest), which usually require high-levels of substantive knowledge and much human effort to obtain (Quinn et al., 2010). In addition, they both have costly prerequisite tasks to perform before the usage. Specifically, dictionary based methods require the building of an appropriate dictionary of words and phrases that are used to indicate the membership in a particular category, and supervised learning methods require a subset of hand-coded texts that will serve as training data for the algorithms.

The motivation of this thesis is therefore to further mitigate the cost of existing automatic text analysis techniques adopted in the financial accounting domain. In particular, we are motivated to provide more effective and efficient solutions for analyzing corporate risk disclosures, which require less (or minimal) cost of human effort to use. In this respect, one of the most promising ideas is the use of the topic models (Blei, 2012), which are algorithms for discovering the main themes (i.e., topics) that pervade a large and otherwise unstructured collection of documents, and do not require any prerequisite manual tasks such as annotations. Refer to (Quinn et al., 2010) for a summary of relative costs associated with major text analysis methods, including dictionary based methods, supervised learning methods, and topic modeling based methods. The role of topic models for text analysis is described below.

1.1.1 The Role of Topic Models

Topic models are a type of statistical model for automatically discovering the abstract "topics" that occur in a collection of documents. The main idea behind the models is the assumption that each document is a mixture of topics, and each topic is a probability distribution over words. To illustrate how topic models work, a real-world example is provided below.

Suppose we are given a collection of seminar abstracts [1]; how can we know: (1)

[1] A collection of 591 seminar abstracts has been collected at the School of Computing, National University of Singapore, from June 2010 to April 2013.

the common themes that pervade these abstracts, and (2) the themes associated with each abstract.

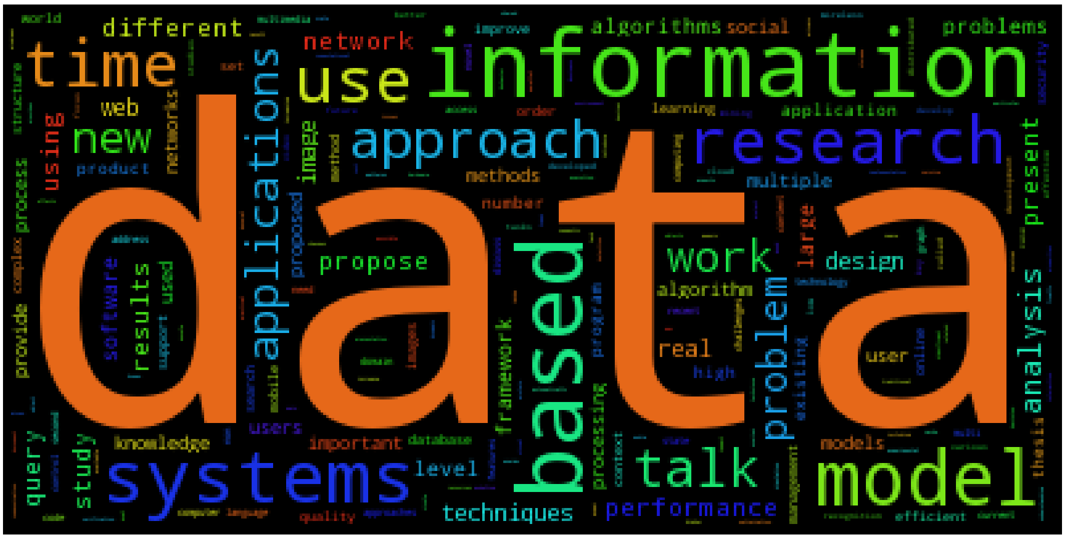
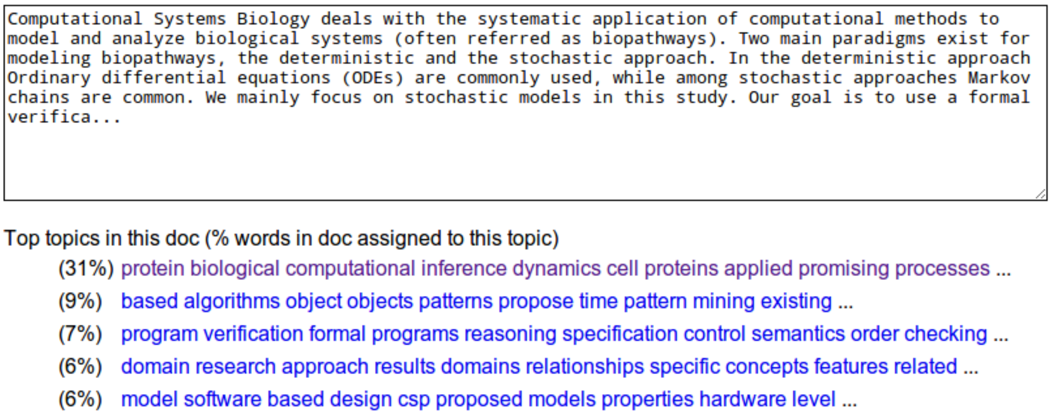


Figure 1.1: Word cloud of the collection-wide frequent words.

The most simple approach might be the manual inspection of each individual abstract, but this is infeasible for large numbers of abstracts due to limited human efforts. A more scalable approach might be to examine the word frequencies in the whole collection. Figure 1.1 shows the word cloud of the top 200 most frequent words after the removal of the non-meaningful stopwords. In a word cloud, the font size is proportional to the frequency in the collection. As can be observed, this method yields few insights into the common themes, since the predominant words are shared across different themes.

A more sophisticated method might be to cluster (Han et al., 2006) each abstract into one cluster, hoping that the clusters could be used to represent the themes. However, almost all abstracts cannot fit into one single cluster (theme). For example, the seminar abstract in Figure 1.2 exhibits different themes (e.g., *computational biology* and *statistical models*), implying that it probably belongs to multiple clusters. Topic models could solve this problem by allowing each document to exhibit multiple topics (themes). The input of the model is a collection of unstructured texts, while the output consists of two components: (1) the collection-wide topics, and (2) the document-wide topic proportions (as well as the topic assignment for each word). Figure 1.2 presents an example of the outputs of topic modeling on the collection of seminar abstracts. At the collection level, we obtain a list of topics in which each topic is represented by a set of words ordered by the corresponding probability [2]. At the document level, we obtain the topic proportions.



```
Computational Systems Biology deals with the systematic application of computational methods to
model and analyze biological systems (often referred as biopathways). Two main paradigms exist for
modeling biopathways, the deterministic and the stochastic approach. In the deterministic approach
Ordinary differential equations (ODEs) are commonly used, while among stochastic approaches Markov
chains are common. We mainly focus on stochastic models in this study. Our goal is to use a formal
verifica...
```

Top topics in this doc (% words in doc assigned to this topic)

(31%) protein biological computational inference dynamics cell proteins applied promising processes ...
(9%) based algorithms object objects patterns propose time pattern mining existing ...
(7%) program verification formal programs reasoning specification control semantics order checking ...
(6%) domain research approach results domains relationships specific concepts features related ...
(6%) model software based design csp proposed models properties hardware level ...

Figure 1.2: An example of the output of topic models.

Topic models discover the latent semantic topics by only looking at the text, i.e., the co-occurrences of words in documents. In particular, the objective of the models is to find the optimal set of latent variables (i.e., topics) that can generate the observed words in documents with maximum likelihood, based on the assumed generative process. However, it is not uncommon that original topic models might discover some topics which appear reasonable but not well-aligned with users' goals in a specific application (Blei, 2012). For example, for the task of extracting risk types (which is one of our research problems that will be introduced later), the direct application of original topic models will lead to the discovery of topics that

[2] We only present a portion of the topics that are discussed in the abstract. The full list of topics is available at http://www.comp.nus.edu.sg/~baoyang/files/nussoc-seminars/all_topics.html

are not meaningful for representing risk types (i.e., the variables of interest). To address this issue, many prior studies propose to extend the original topic models by incorporating additional information accompanied with text, such as document labels, document sources, sentence structures, and so on. The intuition is that textual documents are not just text, and the inclusion of appropriate additional information could steer the model towards topics that are better aligned with the user's goals in the different contexts. Section 2.2 provides a review of such extended topic models. In this thesis, we continue this line of research on extending topic models with the purpose of providing solutions with low costs to be used for analyzing corporate risk disclosures.

1.2 Research Problems

In this thesis, there are two primary goals of analyzing corporate risk disclosures: (1) to interpret the "overall risk sentiment" embodied in the disclosure section, and (2) to extract the various individual risk types enumerated in the section. Clearly, both of these objectives can be reduced to the solving of text analysis problems. We also conduct an empirical study to examine the market reactions to the individual risk types extracted by our proposed solution. The research problems to be investigated in the thesis are elaborated below.

1.2.1 Cross-Industry Risk Sentiment Analysis

Our first goal is to interpret the "overall risk sentiment" embodied in the disclosure section in corporate reports, in particular the "Management Discussion and Analysis" (MD&A) section in 10-K forms. This risk sentiment analysis (Loughran and McDonald, 2011; Li, 2010a) aims to identify the managers' tone in their risk disclosures (either positive or negative), and can be seen as the application of the well-known sentiment analysis (Pang and Lee, 2008) in the financial accounting domain. To solve this problem, there are two types of existing methods, namely the dictionary based methods, and the supervised learning methods. Although the majority of existing works rely on dictionary based methods (Loughran and McDonald, 2011), supervised learning methods have been demonstrated to be more effective in recent studies (Li, 2010a). However, it has been recently shown that sentiment classification (i.e., the supervised learning method) is highly sensitive to the domain from which the training data is annotated (Pang and Lee, 2008; Liu and Zhang, 2012). Specifically, a sentiment classifier trained using opinionated documents from one domain (i.e., the source domain) usually performs poorly when it is applied to opinionated documents from another domain (i.e., the target domain). The reason for this is that word patterns used in different domains to express sentiment can be quite different. In our case, we also suffer from this problem since firms from different industries (i.e., domains) often use industry-specific words to express the risk sentiment. For example, when describing the "product approval" risk, firms in the airline industry tend to use the words like "flying test" while those in the pharmaceutical industry are likely to use the words like "clinical trial". To address this issue, the most straightforward solution is to manually annotate sufficient training data in the target domain, and then apply the conventional supervised learning methods. However, the manual annotation of training data can be quite time-consuming and expensive.

To bridge this gap, in this thesis, we study the problem of cross-industry sentiment analysis, where the risk sentiment classifier is trained using the annotated training documents from one industry (i.e., the source industry), but is meant to be applied in another industry (i.e., the target industry). In particular, our goal is to train a robust sentiment classifier by reusing the available annotated data in the source industry. This classifier should be able to mitigate the word pattern difference between the source and target industries, and accurately predict the risk sentiment in the target industry.

1.2.2 Extracting Individual Risk Types

Our second goal is to extract various individual risk types from the textual disclosures, in particular the “Risk Factors” section (i.e., item 1A) in 10-K forms. At a high level, risk types refer to general factors that present elements of risk to a corporation, such as litigation, human resources, catastrophe, and so on. To solve this problem, all existing methods, namely the dictionary based methods and the supervised learning methods that will be reviewed in Chapter 2 later, assume a pre-defined set of categories (i.e., risk types). 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 tone of textual statements, 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. Unfortunately, all prior work is based on the notion of pre-defined risk types. 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 thesis, we study the problem of extracting individual risk types without pre-defining them. Specifically, our goal is to estimate rather than pre-define a set of categories (i.e., risk types), and simultaneously assign sentences to those categories.

1.2.3 Market Reactions to Individual Risk Types

In addition to the two primary goals aforementioned, we conduct a follow-up study to examine the market reactions to the individual risk types in textual disclosures. In prior literature, there are competing arguments on whether and how risk disclosures affect the risk perceptions of investors. The first argument is that risk disclosures are by and large boilerplate, and therefore have no impact on investors (Schrand and Elliott, 1998). One the other hand, there are also empirical findings that risk disclosures are informative. In particular, some studies suggest that investors' risk perceptions will increase with more risk disclosures, while the others suggest that investors' risk perceptions will decrease. Recently, Kothari et al. (2009) argue that previous mixed findings are due to the tone of risk disclosures. Specifically, favorable disclosures will decrease investors' risk perceptions while unfavorable disclosures will increase them.

Different from prior studies, in this thesis, we hypothesize that the effects of risk disclosures will depend on the their semantic content, i.e., the specific risk types disclosed. In particular, we conduct an empirical study to examine whether and how individual risk types extracted from textual risk disclosures will affect the post-disclosure risk perceptions of investors.

1.3 Contributions

The main contributions of this thesis are summarized as follows.

First, we propose an extended topic model, called PSCCLDA, and its learning algorithm for cross-industry risk sentiment analysis of corporate risk disclosures. As far as we know, this is the first work that introduces the cross-domain learning (also called transfer learning) (Pan and Yang, 2010) into the field of financial accounting. Our model could mitigate the distributional difference between the source and target industries in the low-level word feature space by learning a new high-level topic feature space in a supervised way. This could reduce the amount of human effort for analyzing risk disclosures by reusing the annotated data available in the source industries. Evaluations are conducted on nine standard testing datasets and one real-world risk disclosure dataset, and the results demonstrate the effectiveness of our proposed method.

Second, we propose an extended topic model, called Sent-LDA, and its learning algorithm for extracting individual risk types (e.g., funding risk, infrastructure risk, etc.) from textual disclosures without pre-defining them. As far as we know, this is the first work that introduces the unsupervised learning into the field of financial accounting. Our model could estimate rather than pre-define a set of categories (i.e., risk types), and simultaneously assign sentences to those categories. This could reduce the amount of human effort substantially when analyzing corporate risk disclosures. Experimental results show that our proposed method outperforms all competing methods, and could discover more meaningful topics that are representative for risk types. We further visualize our learned model in a publicly available system [3]. The system facilitates the navigation of large amount of textual risk disclosures by our target user-base, including financial analysts, business managers or academic researchers.

Third, our empirical study on market reactions to individual risk types contributes to the literature on examining the effects of risk disclosures. This is the first study that examines the effects of risk disclosures at the individual risk type level.

[3] The system is available at http://www.comp.nus.edu.sg/~baoyang/10kslda/browse/topic-list.html

Different from prior studies, our results provide support for all three competing arguments regarding whether and how risk disclosures (in section 1A of 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, while the other five types of unsystematic risks will decrease them. Our findings reconcile the conflicted arguments on the effects of risk disclosures, and have implications for both researchers and practitioners in the field of financial accounting.

1.4 Thesis Organization

The rest of this thesis is organized as follows.

In Chapter 2, we review the literature on related works. We begin with the introduction of topic models and its learning algorithms, as well as several related probabilistic models of text. Then, we review the extended topic models in previous works, including supervised topic models, cross-collection topic models, and topic models incorporated with sentence structure. Finally, we review the existing methods for text analysis, including the dictionary based and supervised learning methods that have been adopted in the financial accounting domain, and the unsupervised learning and cross-domain learning methods that have not been adopted yet.

In Chapter 3, we present our solution for cross-industry risk sentiment analysis. We begin with an overview of the study, and then provide the problem formulation.

Next, we describe our proposed model, called PSCCLDA (Partially Supervised Cross-Collection LDA), and its learning algorithm. Finally, we show the experimental evaluation of our model for both the general task of cross-domain text classification, and the specific task of cross-industry risk sentiment analysis in the financial accounting domain.

In Chapter 4, we present our solution for extracting individual risk types from textual disclosures without pre-defining them. We begin with an overview of the study, and then provide the problem formulation. Next, we describe our proposed model, called Sent-LDA (Sentence-based LDA), the intuition behind the model, and its learning algorithm. After that, we show the experimental evaluation of our model for the task of risk type extraction from textual disclosures. Finally, we demonstrate a browser of textual risk disclosures, which visualizes the outputs of our learned model.

In Chapter 5, we continue the analysis of the extracted risk types in Chapter 4. Specifically, we conduct an empirical study to investigate whether and how individual risk types in risk disclosures will affect the post-disclosure risk perceptions of investors. We begin with an overview of the study, introducing the conflicted findings in previous studies. Next, we present the research question and our hypothesis, and then describe the data preparation. After that, we elaborate our econometric model, including the model specification and estimation results. Finally, we discuss the main findings and the implications of our study.

In Chapter 6, we conclude the thesis by providing concluding remarks, limitations, and possible directions for future work.

Chapter 2

Literature Review

In this chapter, we provide a review of previous works that are related to this thesis. We first introduce the topic models, including related probabilistic models of text, LDA topic model, and its learning algorithms. We then review some extended topic models that are closely related to our studies. Finally, we review the existing methods for text analysis, including the dictionary based and supervised learning methods that have been adopted in the financial accounting domain, and the unsupervised learning and cross-domain learning methods that have not been adopted yet.

2.1 Topic Models

There are various types of probabilistic models of text, which have been widely used in the fields of text miming, natural language processing and information retrieval (Sun et al., 2012). For example, NB (Naive Bayes) classifier (McCallum et al., 1998), perhaps the simplest statistical model for classification, has been widely applied in text mining. HMM (Hidden Markov Model) (Rabiner, 1989), which is a powerful statistical model for modeling sequential or time-series data, has been successfully used in many text-related tasks such as the part-of-speech tagging (Kupiec, 1992) in NLP (natural language processing). CRF (Conditional Random Fields) (Lafferty et al., 2001), which is another probabilistic model for sequential data, has been proven to be superior than the HMM model for labeling or segmenting sequential data in many NLP tasks such as part-of-speech tagging (Lafferty et al., 2001) and shallow parsing (Sha and Pereira, 2003). N-gram language model (Ponte and Croft, 1998), which assigns a probability to a sequence of words by means of a probability distribution, has been widely applied in information retrieval.

While there are many types of probabilistic models of text, this thesis focuses on one of them, called topic model, which aims to uncover the underlying semantic structures in unstructured texts. The idea of the topic model is to treat a document as the mixture of topics where a topic is a probability distribution over words. The first milestone of topic model is the PLSA (Probabilistic Latent Semantic Analysis) model proposed by Hofmann (1999a). Later, the LDA (Latent Dirichlet Allocation) model is proposed by Blei et al. (2003), which generalizes the PLSA by casting a generative Bayesian framework to avoid the over-fitting issue suffered by the PLSA (Blei et al., 2003). The main advantage of formulating the LDA as a generative model is that it can be easily extended for discovering topics that are better aligned with the users' goals in different contexts (Blei, 2012). Due to this advantage, we mainly focus on LDA topic model in this thesis. Specifically, we aim to extend the original LDA model for better analyzing corporate risk disclosures.

2.1.1 LDA Topic Model and Related Probabilistic Models

To better understand how topic models work, we first describe the high-level framework of topic modeling technique defined by Blei (2012). This framework contains four key components, including observed data, model assumption, inference algorithm and discovered structure.

- Observed data. Observed data is simply a collection of documents, each of which is a set of discrete words. It should be noted that the topic models are not limited to model textual data, and can be applied for modeling other types of discrete data such as images represented by bag-of-features (Li and Perona, 2005).

- Model assumption. Model assumption of topic models can be thought of as a story in statistical language about how the observed data (i.e., a collection of documents) is generated. Different models (e.g., PLSA and LDA) make different assumptions, which are usually described using the graphical models as will be shown in Figure 2.3 and 2.4 later.

- Inference algorithm. Inference algorithms for topic models are used to estimate the model parameters and latent variables defined in the model assumption. Two most widely used inference algorithms are collapsed Gibbs sampling (Griffiths and Steyvers, 2004) and variational method (Blei et al., 2003), which will be reviewed in Section 2.1.2.

- Discovered structure. As previously described in Section 1.1.1, there are two components of the discovered structure: (1) the inferred topics at the collection level, and (2) the topic proportions at the document level. There are different ways to present the discovered structure, and visualization techniques are usually used for improving the representation (Chuang et al., 2012).

In the following, we introduce the LDA topic model, and some related probabilistic models of text, including the unigram model, the mixture of unigrams, and the PLSA topic model.

Unigram Model

The unigram model is the simplest language model which assumes that a word sequence is generated by sampling each word independently. The graphical representation of unigram model is shown in Figure 2.1. As can be seen, there are no latent variables in the model, and each word w is sampled independently from some distribution. When modeling the discrete textual data, the multinomial distribution is usually used for unigram model, which in turn serves as the component for more complicated mixture models such as topic models (Nigam et al., 2000; Blei et al., 2003). Formally, let M be the number of documents in the corpus, N be the total number of words in a document, V be the set of words in the vocabulary, the probability of the word sequence $w_1, w_2, ..., w_N$ $(w_i \in V)$ in a document d is defined as:

$$p(d) = \prod_{i=1}^{N} p(w_i)$$

where $p(w_i)$ follows a multinomial distribution $Multinomial(\boldsymbol{\theta})$ over all words in the vocabulary.

Obviously, the unigram model makes the strict assumption that all words are generated independently. To capture the dependency between words, the unigram model can be generalized to the so-called n-gram model in which the occurrence of a word depends on the preceding $n-1$ words. Despite its simplicity, the unigram model has been demonstrated to be quite effective for various tasks like information retrieval, while the more sophisticated n-gram models tend not to improve much over it (Zhai, 2008).

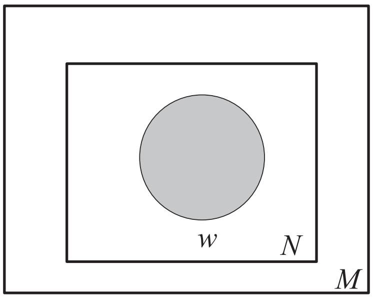


Figure 2.1: Graphical representation of unigram model.

Mixture of Unigrams

The unigram model assumes that all words in the collection are generated from one single multimonial distribution. This implies that it only allows one topic in the whole collection of documents, if we treat the multinomial distribution as a "topic". The mixture of unigrams model (Nigam et al., 2000) relaxes this unrealistic assumption by assuming that there are multiple topics in the collection of documents. More specifically, it augments the unigram model with a discrete latent topic variable z, as shown in the graphical representation in Figure 2.2. In the model, each document is generated by first choosing a topic z, and then generating the N words in the document independently from the conditional multinomial distribution $p(w|z)$. Formally, the probability of the word sequence $w_1, w_2, ..., w_N$ $(w_i \in V)$ in a document d is defined as:

$$p(d) = \sum_{z \in \boldsymbol{Z}} p(z) \prod_{i=1}^{N} p(w_i|z)$$

where $p(z)$ is a multinomial distribution over a fixed set of topics $\boldsymbol{Z}$.

Although the mixture of unigrams model assumes the multiple topics in the text collection, it is still restrictive since it only allows one topic for each document. The breakthrough idea of topic models is to allow each document to exhibit multiple topics. This idea leads to two milestones in topic modeling, i.e., the PLSA (Probabilistic Latent Semantic Analysis) and the LDA (Latent Dirichlet Allocation).

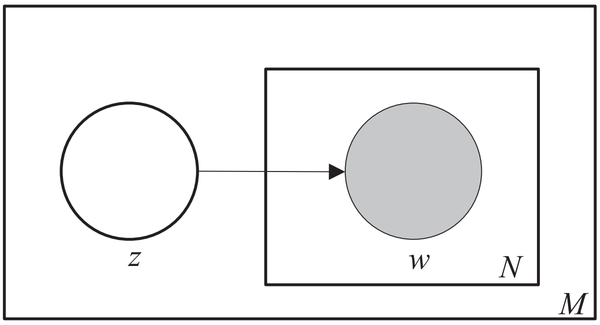


Figure 2.2: Graphical representation of mixture of unigrams.

Probabilistic Latent Semantic Analysis

The PLSA (Probabilistic Latent Semantic Analysis) topic model was established in three papers (Hofmann, 1999a,b, 2001). It is also called PLSI (Probabilistic Latent Semantic Indexing), especially by researchers in the field of information retrieval.

The graphical model of PLSA is given in Figure 2.3. As can be seen, each document is represented as a mixture of $|Z|$ latent topics, and each latent topic z is represented as a multinomial distribution over words $p(w|z)$ in the corpus. To generate a word w in a document d, a topic z is first generated from the document-specific mixture of topics $p(z|d)$, and then the word is generated using the multinomial distribution associated with that topic. Thus, each word is generated from a single topic and words in the same document can be generated by multiple topics. The PLSA is, therefore, more flexible than the aforementioned mixture of unigrams or the cluster model (Dhillon and Modha, 2001) which constrains all words in a document to be associated with a single topic. Formally, the joint probability of an observed document (d, w) is defined as:

$$p(d, w) = p(d)p(w|d) = p(d)\sum_z p(w|z)p(z|d)$$

The PLSA relaxes the simplifying assumption made in the mixture of unigrams model that each document is generated from only one topic. Specifically, it allows each document to exhibit multiple topics, and $p(z|d)$ can serve as the mixture weights of the topics for a particular document d. However, the PLSA is not a well-defined generative model of documents and cannot be naturally applied to a previous unseen document. This is because d is a dummy index into the list of documents in the training set, and the topic proportion $p(z|d)$ is only estimated for those training documents. Besides, the number of model parameters (i.e., $p(z|d)$) will grow linearly with the number of training documents. making the model prone to the over-fitting issue.

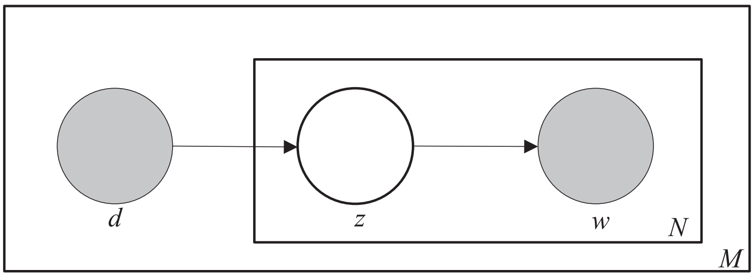


Figure 2.3: Graphical representation of PLSA model.

Latent Dirichlet Allocation

To overcome the limitations of the PLSA, Blei et al. (2003) proposed the LDA (Latent Dirichlet Allocation) topic model which treats the topic mixture weights as a $|Z|$-parameter hidden random variable (Dirichlet distribution) rather than a large set of individual parameters which are explicitly linked to the training documents. Girolami and Kabán (2003) showed that the PLSA is a MAP (Maximum A Posteriori) estimated LDA model under a uniform Dirichlet prior, and therefore the perceived shortcomings of the PLSA can be resolved and elucidated within the LDA framework.

The graphical model of LDA is given in Figure 2.4. As can be seen, the LDA assumes that each document is associated with a multinomial θ over topics, where each topic z is associated with a multinomial β over the words in the vocabulary. To generate a word, a topic z is first chosen according to the topic proportion θ, and then the word is picked based on the chosen topic z. To complete the model, the Dirichlet prior α and η is placed over θ and β respectively. These priors are chosen due to the conjugation between the multinomial and Dirichlet distribution (Blei et al., 2003), which could result in the computational convenience.

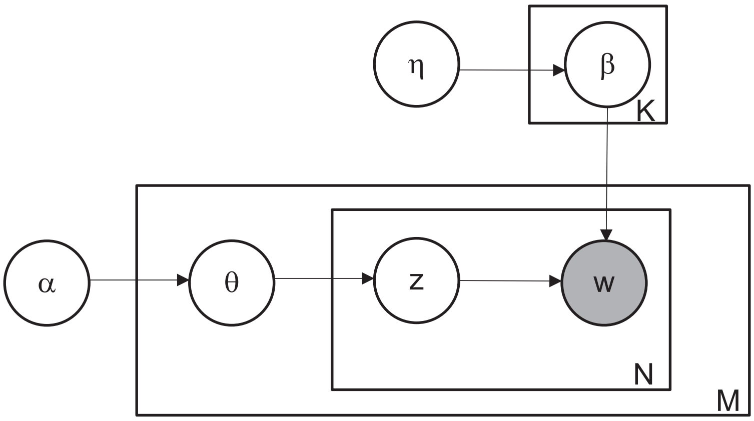


Figure 2.4: Graphical representation of LDA model.

Formally, let M, N, K, 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. $Dirichlet(\cdot)$ and $Multinomial(\cdot)$ are Dirichlet and Multinomial distribution with parameter $(\cdot)$ respectively. $\boldsymbol{\beta_k}$ is the V-dimensional word distribution for topic k, and $\boldsymbol{\theta_d}$ is the K-dimensional topic proportion for document d. $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$ are the hyper-parameters of the corresponding Dirichlet distributions. The graphical representation of LDA is shown in Figure 2.4, and the corresponding generative process is:

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

This generative process implies the joint probability of the observed documents (w) and hidden variables (z, θ, β) as follows:

$$p(w, z, \theta, \beta) = \left(\prod_d p(\theta|\alpha)\right)\left(\prod_z p(\beta|\eta)\right)\left(\prod_{w_i} p(w_i|\beta_{z_i})p(z_i|\theta_{d_i})\right)$$

2.1.2 Learning Algorithms for LDA Model

We now introduce the learning algorithms for the LDA topic model. The key inferential problem for learning the LDA model is that of computing the posterior distribution of the hidden variables (i.e., topic assignments z for words and topic proportions θ for documents) given the model parameters (i.e., topic distributions β and hyper-parameters) and the observed documents (i.e., observed words w):

$$p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\boldsymbol{w})} \tag{2.1}$$

Unfortunately, this distribution is intractable to compute in general (Blei et al., 2003). The learning algorithms for topic models usually approximate the Equation 2.1 by forming an alternative distribution over the latent topic structure that is adapted to be close to the true posterior. These algorithms generally fall into two categories (Blei, 2012): *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 is the *collapsed Gibbs sampling* method proposed by Griffiths and Steyvers (2004). An alternative to the sampling based algorithms are the *variational methods*. Rather than approximating the posterior with samples, variational methods posit a parametrized 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 EM (Expectation Maximization) algorithm proposed by Blei et al. (2003).

In the following, we describe some details of both the collapsed Gibbs sampling and variational EM algorithms for learning the LDA model. These two algorithms form the basis for the derivation of learning algorithms of our proposed models.

Collapsed Gibbs Sampling

There are two parameters of the LDA to be estimated, namely the topic probability over terms ϕ [1] and the topic proportions of document θ. Gibbs sampling is an effective strategy for estimating these two parameters. It is an approximate iterative technique which is a special form of Markov Chain Monte Carlo (MCMC) (Bishop and Nasrabadi, 2006).

Rather than explicitly estimate ϕ and θ, Gibbs sampling method approximates the

[1] For convenience, we abuse the notation here. For variational methods, the topic-word distribution is usually denoted as β such as in (Blei et al., 2003). But for Gibbs sampling methods, ϕ is more often used to denote the topic-word distribution while β is used to denote its hyper-parameters.

posterior distribution of topics given observed words $p(z|w)$ by means of Monte Carlo algorithm. Specifically, it iterates over each word token in the text collection in a random order and estimates the probability $p(z_i = k)$ of assigning the current word token i to a topic k, conditioned on the topic assignments to all other word tokens z_{-i} as follows:

$$P(z_i = k|z_{-i}, w, \alpha, \beta) \propto \left(\frac{n^{d_i}_{-i,k} + \alpha}{n^{d_i}_{-i,\cdot} + T\alpha} \right) \left(\frac{n^{w_i}_{-i,k} + \beta}{n^{(\cdot)}_{-i,k} + W\beta} \right) \tag{2.2}$$

where $n^{d_i}_{-i,k}$ is the number of times that topic k is assigned to some words in document d_i, not including the current instance i; $n^{w_i}_{-i,k}$ is the number of times that word w_i is assigned to topic k, not including the current instance i; T is the number of topics; W is the number of distinct words in the vocabulary; α and β are the symmetrical hyper-parameters for the document-topic and topic-word Dirichlet distributions, respectively. A missing subscript or superscript (e.g., $n^{(\cdot)}_{-i,k}$) indicates a summation over that dimension. $-i$ indicates that the counts are calculated by omitting the current instance i.

In each iteration, a topic is sampled, based on Equation 2.2, for each word in the collection. After sufficient iterations, the sample obtained can be used to approximate the model parameters. Specifically, the topic-word distribution ϕ and the document-topic distribution θ can be estimated as follows:

$$\phi^w_k = \frac{n^w_k + \beta}{n^{(\cdot)}_k + W\beta} \tag{2.3}$$

$$\theta^k_d = \frac{n^k_d + \alpha}{n^{(\cdot)}_d + T\alpha} \tag{2.4}$$

The Gibbs sampling method for the LDA described here is first proposed by

Griffiths and Steyvers (2004). More precisely, it is called collpased Gibbs sampling due to the fact that θ and ϕ are integrated out when deriving Equation 2.2. The detailed derivation of the above equations can be found in (Heinrich, 2005).

Variational Methods

Variational methods are another type of learning algorithms that have been successfully applied to many kinds of topic models, where corpus size and vocabulary dimension are large (Wainwright and Jordan, 2008). The basic idea of variational methods is to introduce a variational distribution $q(\theta, z|\gamma, \phi)$ to approximate the intractable posterior distribution $p(\theta, z|w, \alpha, \beta)$ in Equation 2.1, where γ and ϕ are variational parameters. To get a tractable variational distribution, Blei et al. (2003) break the coupling between θ and β as shown in Figure 2.4, and define the variational distribution as the following factorized form:

$$q(\theta, z|\gamma, \phi) = q(\theta|\gamma)q(z|\phi) \tag{2.5}$$

where $q(\theta|\gamma)$ follows a Dirichlet distribution and $q(z|\phi)$ follows a Multinomial distribution.

The next step is to formally specify an optimization problem to determine the values of γ and ϕ. In particular, Blei et al. (2003) show that finding an optimal lower bound on the log likelihood results in the following optimization problem:

$$q(\gamma^*, \phi^*) = argmin_{(\gamma,\phi)} D(q(\theta, z|\gamma, \phi)||p(\theta, z|w, \alpha, \beta)) \tag{2.6}$$

which is a minimization of the Kullback-Leibler (KL) divergence (Bishop and Nasrabadi, 2006) between the variational distribution and the actual posterior distribution. One method to minimize this function is to use an iterative fixed-point method (Blei et al., 2003), yielding update equations of:

$$\phi_{ni} \propto \beta_{iw_n} exp\left\{E_q[log(\theta_i)|\gamma]\right\} \tag{2.7}$$

$$\gamma_i = \alpha_i + \sum_{i=1}^{N} \phi_{ni} \tag{2.8}$$

Here, the expectation in the ϕ update in Equation 2.7 is computed as:

$$E_q[log(\theta_i)|\gamma] = \Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j) \tag{2.9}$$

where Ψ is the first derivative of the log gamma function, which can be computed via a Taylor approximation.

It is worth mentioning that Equations 2.7 and 2.8 have an appealing intuitive interpretation. The Dirichlet update is a posterior Dirichlet given expected observations taken under the variational distribution, $E[z_n|\phi_n]$. The multinomial update is akin to using Bayes theorem, i.e., $p(z_n|w_n) \propto p(w_n|z_n)p(z_n)$, where $p(z_n)$ is approximated by the exponential of the expected value of its logarithm under the variational distribution, and $p(w_n|z_n) = \beta_{z_n w_n}$.

The variational inference described above assumes that we have known the topic-word distribution β, and the hyper-parameter α. But how to find β and α? To tackle this problem, Blei et al. (2003) propose the so-called variational EM algorithm which uses the EM (Expectation Maximization) algorithm with variational distribution. Specifically, we can find the empirical Bayes estimates for the LDA model via an alternating variational EM procedure which maximizes a lower bound with respect to the variational parameters γ and ϕ, and then maximizes the lower bound with respect to the model parameters α and β by fixing the values of the estimated variational parameters.

The detailed derivation of the variational EM algorithm for LDA can be found in (Blei et al., 2003). The derivation yields the following iterative algorithm:

- E-step: For each document, find the optimized values of the variational parameter $\{\gamma_d^*, \phi_d^* : d \in D\}$. This is done by variational inference described previously in this section.
- M-step: Maximize the resulting lower bound on the log likelihood with respect to the model parameters α and β. The objective is to find the maximum likelihood estimates with expected sufficient statistics computed in the E-step.

These two steps are iterated alternately until the lower bound on the log likelihood converges.

It should be noted that we can choose to update the hyper-parameter α using Newton-Raphson method as in (Blei et al., 2003), or keep it fixed during the EM iterations. But for the collapsed Gibbs sampling method, all the hyper-parameters are fixed and have to be pre-set empirically.

2.2 Extended Topic Models

Original topic models (i.e., PLSA and LDA) discover the latent semantic topics by only looking at the text, i.e., the co-occurrences of words in documents. However, it is not uncommon that original topic models might discover some topics which are reasonable but not well-aligned with the users' goal in a specific application

(Blei, 2012). To address this issue, many prior studies propose to extend the original topic models by incorporating the additional information accompanied with text. The intuition is that the documents are not just text, and the inclusion of appropriate information could steer the model towards topics that are better aligned with the user's goals in different contexts.

In the following, we review three types of extended topic models that are closely related to our studies. In particular, the supervised topic models and the cross-collection topic models are related to the study in Chapter 3, and the topic models incorporated with sentence structure are related to the study in Chapter 4.

2.2.1 Supervised Topic Models

Original topic models (i.e., PLSA and LDA) are unsupervised models which could discover the broad patterns (i.e., topics) in a document collection without any supervision. However, they are usually inadequate for the prediction tasks, such as document classification and sentiment analysis (Blei and McAuliffe, 2008). Consider the task of predicting a movie rating from its associated review text. Intuitively, good predictive topics should be able to distinguish words like "excellent", "terrible", and "average", without regard to genre. But it is possible that the topics estimated by an unsupervised topic model will correspond to genres, if that is the dominant thematic structure in the text collection.

To address this issue, some supervised topic models are recently proposed with the purpose of enhancing topic models for prediction tasks. The basic idea is to incorporate the associated labels to be predicted into the generative process of the extended models. This allows the model to explicitly explore the associations between labels and topic features so that more predictive topics can be discovered.

Examples of such models include the supervised LDA (Wang et al., 2009a; Blei and McAuliffe, 2008), the labeled LDA (Ramage et al., 2009), the partially labeled LDA (Ramage et al., 2011), the DiscLDA (Lacoste-Julien et al., 2008), the Dirichlet-multinomial regression model (Mimno and McCallum, 2008), and the MedLDA (Zhu et al., 2009).

2.2.2 Cross-Collection Topic Models

Original topic models have been demonstrated to be effective in modeling a single collection of textual documents. However, as indicated in (Zhai et al., 2004), they are inadequate for modeling text from different collections for two reasons. First, the structure of collections will be completely ignored, and the extracted topics might only represent some but not all collections. Second, it is hard to identify whether a topic represents the common information across collections or the specific information of a particular collection.

To bridge this gap, two variants of topic model are proposed to model multiple text collections. The first variant, called CCMix (Cross-Collection Mixture), is proposed by Zhai et al. (2004). It extends the PLSA model by explicitly distinguishing common topics that characterize the common information across all collections from specific topics that characterize the collection-specific information. Common topics and collection-specific topics are aligned under the same set of indices and the number of topics in each collection are forced to be the number of common topics. The second variant, called CCLDA (Cross-Collection Latent Dirichlet Allocation), is proposed by Paul and Girju (2009). It further extends the CCMix by replacing the PLSA framework with that of the LDA. Thus, it is actually a Bayesian version of the CCMix model.

The cross-collection topic models have been successfully applied for many problems. The most direct application is the comparative text analysis. For example, they have been used for the comparative text analysis of news articles about wars (Zhai et al., 2004) and the cross-cultural analysis of blog articles (Paul and Girju, 2009). In addition, they have been extended for modeling multiple text streams with temporal dynamics (Hong et al., 2011).

It is also natural to extend the cross-collection topic models for the task of cross-domain text classification (which will be reviewed in Section 2.3.4 later), if we treat documents in each domain as a collection. However, as far as we know, there are no such attempts previously. In Chapter 3, we will further extend the CCLDA model to a supervised version which could be directly applied for the cross-domain text classification in general, and the cross-industry risk sentiment analysis in particular.

2.2.3 Topic Models Incorporated with Sentence Structure

Original topic models make the "bag-of-words" assumption, which states that the order of words in a document does not matter. This simplified assumption is clearly unrealistic, since the change of word order will probably result in the change of meaning expressed in a sentence or a document. Therefore, some researchers have recently proposed to relax or modify the unrealistic "bag-of-words" assumption for uncovering more meaningful topics. The basic idea is to incorporate the information of word order into the model. For example, Griffiths et al. (2005) proposed a model which considers the order of a sequence of words by allowing the model to switch between the LDA model and the standard HMM (Hidden Markov Model) model. Wallach (2006) proposed to combine the LDA model and the bigram model by assuming that the generation of a word is conditioned on both the topic and the preceding word. Blei and Lafferty (2006) proposed the dynamic topic model which considers the temporal order of documents by modeling a topic as a sequence of distributions over words rather than a single distribution.

In this direction, our work in Chapter 4 is closely related to the topic models incorporated with sentence structure. Specifically, there are some recently proposed models which exploit the sentence structure for modifying the "bag-of-words" assumption. Distinct from our proposed "one topic per sentence" assumption in Chapter 4, 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., 2009b; 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 the Local-LDA (Brody and Elhadad, 2010) performs best despite its simplicity.

2.3 Existing Methods for Text Analysis

This thesis is closely related to the well-known research area on automated text analysis, which aims to quantify textual information into numeric variables of interest. As surveyed in (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. In this section, we mainly focus on the application of automatic text analysis in the field of financial accounting.

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. Take the corporate risk disclosure for example. Many financial accounting researchers are interested in the tone (either positive or negative) and the types of risks (e.g., potential lawsuits, catastrophes, etc.) contained in the textual disclosures. In this scenario, the goal is to categorize each unit (e.g., word or sentence) of disclosure documents into one or more categories (i.e., positive or negative sentiment, or various 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 reducing the amount of human effort.

In this section, we review four types of existing methods for automated text analysis, including: (1) dictionary based methods, (2) supervised learning methods, (3) unsupervised learning methods, and (4) cross-domain learning methods. The first two types of methods have been successfully applied in the financial accounting domain, while the latter two types of methods have not been adopted yet.

2.3.1 Dictionary Based Methods

Dictionary based methods are the most simple and intuitive automated text categorization methods. The idea is to first build a dictionary of key words or phrases for indicating the membership of categories. Once the dictionary is built, it can be used to classify documents into categories or measure the extent to which documents belong to a particular category.

Due to its simplicity, dictionary based methods have been widely adopted for text analysis in financial accounting research. For example, they have been used for measuring the tone and sentiment of textual disclosures, such as corporate annual reports (Kothari et al., 2009; Feldman et al., 2010; Kravet and Muslu, 2013; Loughran and McDonald, 2011), news articles (Tetlock, 2007; Tetlock et al., 2008), earning announcements (Rogers et al., 2011), investor message boards (Antweiler and Frank, 2004), Initial Public Offering (IPO) prospectus (Loughran and McDonald, 2013), and so on.

To build the dictionary for sentiment analysis, a commonly used source for word classification is the Harvard Psychosociological Dictionary, particularly the Harvard-IV-4-TagNeg (H4N) file. Recently, Loughran and McDonald (2011) showed that word lists created for other disciplines (i.e., H4N file) misclassified common words in financial text, and developed an alternative dictionary that better reflects the tone in financial context.

Apart from the *tone* (sentiment), there are some other variables of interest to quantify in financial disclosures, including *amount*, *readability* and *risk type* (Li, 2010b). Take the *risk type*, which is most related to this thesis, for example. Campbell et al. (2014) created a keyword list by risk category based on the dictionaries used in prior works and then used this list for classifying risk disclosures in section 1A of 10-K forms into five categories, including systematic, idiosyncratic, financial, tax and legal risks.

Dictionary based 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 in which they were originally developed.

2.3.2 Supervised Learning Methods

Supervised learning methods provide an alternative method for assigning documents (or other units of analysis) to pre-defined categories. The idea is that: (1) human coders first categorize a set of documents by hand; (2) the algorithm then learns how to assign categories to documents using coded data (training set). Supervised learning methods have two major advantages over dictionary based 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 been successfully applied for text analysis in social science research (Hopkins and King, 2010), and have been recently introduced into the field of financial accounting for analyzing textual disclosures. For example, Li (2010a) used a Naive Bayesian classifier, one of the most popular supervised learning methods, to classify the tone and content of forward-looking statements in corporate 10-K and 10-Q filings. Huang and Li (2011) developed a multi-label text classification algorithm to classify risk disclosures in "Risk Factor" section of 10-K form into 25 risk types. Humpherys et al. (2011) proposed to use linguistic features to distinguish fraudulent from non-fraudulent 10-K reports using off-the-shelf supervised classifiers. Cecchini et al. (2010) developed a method for automatically creating an ontology for texts in MD&A section of 10-K forms, which could then be used for classifying financial events of firms.

2.3.3 Unsupervised Learning Methods

Dictionary and supervised learning methods assume a pre-defined 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 it requires a single, precisely defined objective function that works across applications. This is infeasible given that human beings are typically optimizing a (mathematically ill-defined) goal of "insightful" or "useful" conceptualizations. In other words, it is not uncommon that unsupervised models yield clusters that do not correspond to what the user had in mind. Grimmer and Stewart (2013) pointed out that there are two strategies to tackle this problem.

- **Strategy 1.** The first strategy is to allow users to efficiently search over the potential categorization schemes for identifying interesting or useful organizations of the texts. For example, Grimmer and King (2011) developed a computer-assisted method for the discovery of insightful conceptualizations in the form of clustering of input objects.

- **Strategy 2.** 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 need the variation of models. For example, Grimmer (2010) proposed a statistical model that attends to the structure of political rhetoric when measuring expressed priorities: statements are naturally organized by author. Their expressed agenda model exploits this structure to simultaneously estimate the topics in the texts, as well as the attention political actors allocate to the estimated topics.

As far as we know, no previous works attempt to use unsupervised clustering methods for categorizing corporate risk disclosures. In Chapter 4, we will 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 dedicated to the estimated topics.

2.3.4 Cross-Domain Learning Methods

Conventional supervised learning algorithms assume that the training and testing data follow the identical distribution. However, in practice we may have a source domain with plentiful labeled training data, but need to classify the unlabeled data from a target domain which has a different distribution from the source domain. Suppose, for example, we want to classify blog articles into some predefined categories (e.g., sports, politics, etc.). There are usually no labeled data in this target domain (i.e., blog articles) but abundant labeled data in another source domain such as news articles which are well-organized in news websites like CNN.com and BBC.com. The data distributions in these two domains might be quite different because of different word usages or writing styles. In this scenario, the performance of supervised learning algorithms will normally drop due to the violated assumption of identical distribution. To tackle this problem, *cross-domain learning* (also called *domain adaptation* or *transfer learning*) methods have been recently proposed (Pan and Yang, 2010).

As surveyed in (Pan and Yang, 2010), there are different settings of cross-domain learning. This thesis focuses on the transductive learning setting – there are no labeled data in the target domain but abundant labeled data in the source domain, and the learning tasks in both domains are the same. The existing methods for transductive cross-domain learning can be roughly categorized into two types (Pan and Yang, 2010), including instance-based methods and feature representation based methods. Here, we only review the feature representation based methods, which are closely related to our work in Chapter 3.

Feature representation based methods aim to induce a common feature representation for reducing the distributional difference between the source and target domains. To this end, one type of algorithms attempts to make use of the domain-independent "pivot features" to align the domain-specific features. For example, Blitzer et al. (2006) proposed the SCL (Structural Correspondence Learning) algorithm to learn a low-dimensional latent feature space by exploring the relationships between "pivot features" and "non-pivot features". Pan et al. (2010) proposed the SFA (Spectral Feature Alignment) algorithm to align domain-specific words from different domains into united clusters, with the help of domain-independent "pivot features". Then the domain-independent and domain-specific features are co-clustered into a common latent space. However, the success of this kind of methods crucially depends on the auxiliary tasks for selecting "pivot features", which can be a non-trivial engineering problem for many different applications. Different from these algorithms, our proposed model in Chapter 3 does not rely on any auxiliary tasks.

Another type of algorithms, which are closely related to our work, seeks to take advantage of the topic modeling technique to induce the high-level topic feature space. For example, Xue et al. (2008) proposed the TPLSA (Topic-bridged PLSA) model which extends the PLSA by introducing the supervision of labeled data in the training domain via the pair-wise constraints. However, the TPLSA does not explicitly model the domain-independent and domain-specific topics and simply assumes that the topics are shared by all domains. Zhuang et al. (2010) proposed the CDPLSA (Collaborative Dual-PLSA) model which extends the Dual-PLSA in (Yoo and Choi, 2009). The Dual-PLSA separately models the word topics and document topics. The CDPLSA further assumes that word topics and document topics are respectively independent of the data domain, while the association between word topics and document topics is stable across domains. Although it is claimed that the word topics in different domains are semantically related to each other, the CDPLSA actually only extracts domain-specific (word) topics. Different from the TPLSA and the CDPLSA, Li et al. (2012) proposed the TCA

(Topic Correlation Analysis) method which explicitly extracts the shared topics and domain-specific topics, and utilizes the correlations between them for cross-domain learning. Their experimental results show that the TAC outperforms the TPLSA and the CDPLSA. However, it requires an additional step to align domain-specific topics and the information of labels in the training domain is not utilized (as supervision) for learning the latent topics.

As far as we know, no previous works attempt to use cross-domain learning methods for analyzing corporate risk disclosures. In Chapter 3, we will report the first work on the risk sentiment analysis of corporate risk disclosures in the context of cross-domain (i.e., cross-industry) learning.

Chapter 3

Cross-Industry Risk Sentiment Analysis

In this chapter, we present an extended LDA model and its learning algorithm for cross-industry sentiment analysis of corporate risk disclosures.

3.1 Overview

Risk sentiment analysis (Loughran and McDonald, 2011; Li, 2010a) aims to identify the managers' tone in their risk disclosures (either positive or negative), and can be seen as the application of the well-known sentiment analysis (Pang and Lee, 2008) in the financial accounting domain. To solve this problem, there are two types of existing methods, namely the dictionary based methods and the supervised learning methods as reviewed in Section 2.3. Although the majority of existing works rely on dictionary based methods (Loughran and McDonald, 2011), supervised learning methods have been demonstrated to be more effective in recent studies (Li, 2010a). However, it has been recently shown that sentiment classification (i.e., supervised learning method) is highly sensitive to the domain from which the training data is annotated (Pang and Lee, 2008; Liu and Zhang, 2012). Specifically, a sentiment classifier trained using opinionated documents from one domain (i.e., the source domain) usually performs poorly when it is applied to opinionated documents from another domain (i.e., the target domain). The reason for this is that word patterns used in different domains to express sentiment can be quite different. In our case, we also suffer from this problem since firms from different industries (i.e., domains) often use industry-specific words to express the risk sentiment. For example, when describing the "product approval" risk, firms in the airline industry tend to use the words like "flying test" while those in the pharmaceutical industry are likely to use the words like "clinical trial". To address this issue, the most straightforward solution is to manually annotate sufficient training data in the target domain, and then apply the conventional supervised learning methods. However, the manual annotation can be quite time-consuming and expensive.

With the purpose of learning more robust classifiers in cases where available annotated training data comes from domains that differ from the target domain (i.e., the domain in which the learned classifier is meant to be applied), *cross-domain learning* (also called *domain adaptation* or *transfer learning*) methods have been recently proposed (Jiang, 2008; Pan and Yang, 2010). As reviewed in Section 2.3.4, one of the most promising ideas is to induce a new feature representation so that the distributional difference between domains can be reduced and a more accurate classifier can be learned in the new feature space. Based on this idea, there are some recent works which attempt to employ topic models (e.g., LDA, PLSA) to transform the original word feature space to a new latent topic feature space. The underlying assumption of these topic modeling based methods is that the induced topic features may bridge the gaps between domains by linking domain-specific word features together. This is illustrated in Figure 3.1. Specifically, it is quite common that different domains tend to use different words (w_{si} and w_{ti}) to describe the same topic (t_i). If w_{si} and w_{ti} can be projected into the same topic t_i shared by domains, the knowledge (e.g., the associations between features and class labels) in the source domain can be transferred to the target domain.

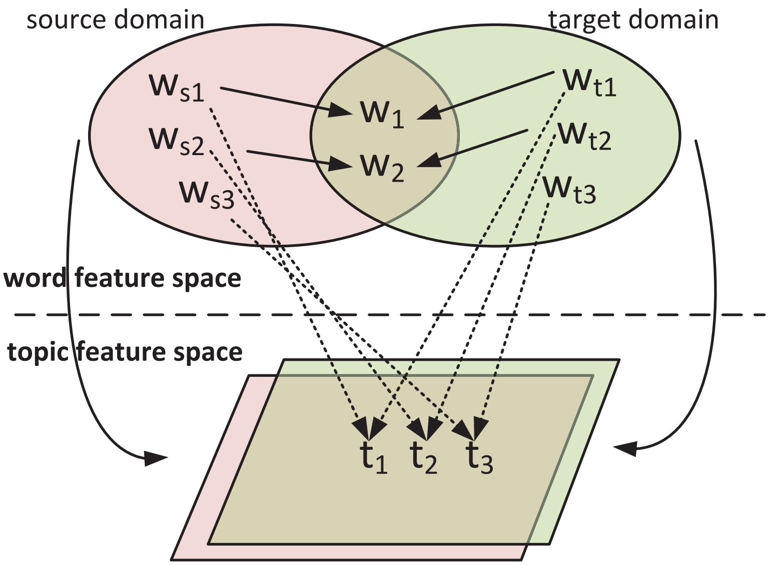


Figure 3.1: Topics as features for bridging domains.

However, we observe two limitations of existing topic modeling based methods for cross-domain learning. First, the label information (e.g., risk sentiment labels) in the training domain cannot be utilized (as supervision) for inducing topic features since the standard topic models are unsupervised. This may lead to the topic features that are not predictive for labels. Specifically, Long et al. (2012) recently demonstrated that the optimal capability of knowledge transfer cannot be achieved if the label information is not used to reinforce the learning of topic features. Second, all induced topics features are assumed to be domain-independent and shared by domains. Unfortunately, this might not be the case since standard topic models can not explicitly distinguish domain-specific and domain-independent topics. Some recent works (Titov, 2011) noticed this issue and proposed to only induce the domain-independent common features that can generalize between domains. More recently, Li et al. (2012) demonstrated that the alignment of domain-specific latent features, in addition to the domain-independent features, can further improve the cross-domain learning. These two limitations are formalized in (Ben-david et al., 2006), which theoretically demonstrated that the designed feature representation should simultaneously minimize the difference between the source and target domains and the empirical training error in the source domain.

To address the two observed limitations above, in this study, we propose an extended LDA topic model for the problem of cross-domain text classification (in general), and cross-industry risk sentiment analysis (in particular). Specifically, our proposed model could (1) explicitly distinguish the domain-independent and domain-specific topics by resorting to the cross-collection topic models (which have been reviewed in Section 2.2.2), and (2) exploit the label information for inferring more predictive topics by embedding the supervised logistic regression model in the similar way as the supervised topic models (which have been reviewed in Section 2.2.1).

The rest of this chapter is organized as follows. Section 3.2 describes the problem formulation. Section 3.3 elaborates our proposed model and its learning algorithm. Section 3.4 presents the experiments for evaluating the proposed model. Finally, Section 3.5 provides a brief summary of the study in this chapter.

3.2 Problem Formulation

In this study, we investigate the problem of cross-industry sentiment analysis, where the risk sentiment classifier is learned using available annotated training data from one industry (i.e., the source industry), but is meant to be applied in another industry (i.e., the target industry). It should be noted that the sentiment analysis is essentially a text classification problem, which aims to classify a document (or other unit of analysis such as sentence or paragraph) into two (binary classification) or more (multi-class classification) sentiment categories (e.g., positive, negative, neutral). Actually, cross-domain sentiment analysis is usually generalized to the cross-domain text classification problem in existing literature (Pan et al., 2010). Similarly, our cross-industry risk sentiment analysis is a special case of cross-domain text classification, and each industry can be seen as a domain. It should also be noted that there are different learning settings for cross-domain text classification (Pan and Yang, 2010). This study particularly focuses on the transductive cross-domain learning setting, in which both the annotated training data in the source domain and the unannotated testing data in the target domain will be utilized for learning the classifier. To sum up, this study investigates the problem of cross-domain text classification in general, and cross-industry risk sentiment analysis in particular [1].

More formally, our problem is defined as follows. Given a source domain (industry) $\mathcal{D}^s = \{(x_1^s, y_1^s), ..., (x_{N_s}^s, y_{N_s}^s)\}$ with N_s labeled documents, and a target domain (industry) $\mathcal{D}^t = \{x_1^t, ..., x_{N_t}^t\}$ with N_t unlabeled documents, our task is to assign a binary class label $y \in \{-1, 1\}$ (-1 denotes the negative risk sentiment label and 1 denotes the positive label) to each unlabeled document x_i^t in the target domain (industry) $\mathcal{D}^t$. We assume that training and testing documents come from related but different domains (industries) $\mathcal{D}^s$ and $\mathcal{D}^t$, and the word feature spaces of the source and target domains (industries) are different, i.e., $\mathcal{X}^s \neq \mathcal{X}^t$.

3.3 Proposed Model

In this section, we elaborate our proposed model and its learning algorithm for cross-domain learning.

[1] For convenience, we sometimes use the term "cross-domain text classification" and "cross-industry risk sentiment analysis" interchangeably in the rest of this chapter.

3.3.1 Model Description

Before describing our model, we first define some notations. We use the convention that lowercase letters denote values (e.g., y, z_{d_i}, x_{d_i}, etc.), and bold letters denote vectors (e.g., $\boldsymbol{x}$, $\boldsymbol{\psi}$, $\boldsymbol{\eta}$, etc.). For convenience, we summary the frequently used notations in Table 3.1.

Table 3.1: Notations for PSCCLDA model.

Symbol	**Description**
$\mathcal{D}^s$, $\mathcal{D}^t$	source/target domain
d_i	word index of document d
z_{d_i}	topic index of word d_i
x_{d_i}	switching variable of word d_i
c_d	collection index of document d
y_d	class label of document d
$\boldsymbol{\phi}_z^C$	common topic z
$\boldsymbol{\phi}_{c,z}^S$	specific topic z of collection c
$\boldsymbol{\psi}_{c,z}$	switching variable distribution
$\boldsymbol{\eta}$	logistic regression coefficients
$\overline{\boldsymbol{z}}_d$	empirical topic distribution for document d
$\boldsymbol{\theta}_d$	topic proportion for document d
$\boldsymbol{\alpha}_c$	collection-specific hyper-parameter for $\boldsymbol{\theta}_d$
$\boldsymbol{\beta}^C$	hyper-parameter for $\boldsymbol{\phi}_z^C$
$\boldsymbol{\beta}^S$	hyper-parameter for $\boldsymbol{\phi}_{c,z}^S$
$\boldsymbol{\gamma}$	hyper-parameter for $\boldsymbol{\psi}_{c,z}$
$Bern(\cdot)$	Bernoulli distribution with parameter $(\cdot)$
$Mult(\cdot)$	Multinomial distribution with parameter $(\cdot)$
$Beta(\cdot)$	Beta distribution with parameter $(\cdot)$
$Dir(\cdot)$	Dirichlet distribution with parameter $(\cdot)$

We now proceed to describe our proposed model, called PSCCLDA (Partially Supervised Cross-Collection LDA). The goal of our model is to induce a new topic feature space in which we can achieve better performance for cross-domain learning. On one hand, it is intuitively necessary to model documents in each domain as a separate collection due to the distributional difference between domains. To this end, we seek to take advantage of the CCLDA (Cross-Collection LDA) (Paul and Girju, 2009) to model documents from multiple collections and explicitly distinguish the collection-independent and collection-specific topics. On the other hand, although the CCLDA could better model documents from multiple collections, it is inadequate for inducing predictive topic features due to its unsupervised learning fashion. We thus propose to further extend the CCLDA to a supervised version for exploiting the label information. We call our model "partially supervised" since we only observe the class labels in the training source domain.

The graphical representation of our model is shown in Figure 3.2, and the corresponding generative process associated is as follows:

1. For each topic z, draw a collection-independent multinomial distribution $\boldsymbol{\phi}_z^C \sim Dir(\boldsymbol{\beta}^C)$
2. For each collection c,
   (a) For each topic z, draw a collection-dependent multinomial distribution $\boldsymbol{\phi}_{c,z}^S \sim Dir(\boldsymbol{\beta}_{c,z}^S)$
   (b) For each topic z, draw a beta distribution $\boldsymbol{\psi}_{c,z} \sim Beta(\boldsymbol{\gamma})$
3. For each document d,
   (a) Draw a topic mixture $\boldsymbol{\theta}_d \sim Dir(\boldsymbol{\alpha}_c)$
   (b) If d is from collection c in source domain $\mathcal{D}^s$, draw a class label $y_d \in \{-1, 1\} \sim Bern(logistic(-y_d \boldsymbol{\eta}^T \overline{\boldsymbol{z}}))$, where $\overline{\boldsymbol{z}}$ is the empirical topic frequencies in d

(c) For each word position i in d,

i. Draw a topic assignment $z_{d_i} \sim Mult(\boldsymbol{\theta}_d)$

ii. Draw a switching variable $x_{d_i} \sim Bern(\boldsymbol{\psi}_{c,z})$

iii. If $x_{d_i} = 0$, draw a word $d_i \sim Mult(\boldsymbol{\phi}_z^C)$;

If $x_{d_i} = 1$, draw a word $d_i \sim Mult(\boldsymbol{\phi}_{c,z}^S)$

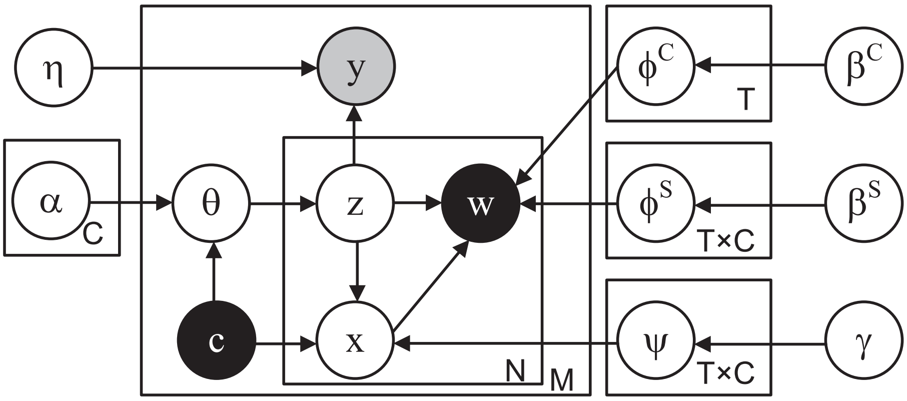


Figure 3.2: Graphical representation of PSCCLDA model.

In our model, there is a set of C collections, and each collection c corresponds to a domain in the context of cross-domain learning. The class labels are observed in the source domain (collection) but not in the target domain (collection). Each collection c is associated with a set T^S of collection-specific topics, and a set T^C of common topics shared by all collections. Similar to the models proposed in (Paul and Girju, 2009; Hong et al., 2011), we assume that there is the same number of elements in all topic sets (i.e., $|T^C| = |T^S| = |Z|$), and the specific topics in different collections are forcibly indexed using the same set of topic ids ($\{1, 2, ..., |Z|\}$) as the common topics. This enables the alignment of the unrelated specific topics in different collections under the same topic index. Thus, the total number of topics in our model is $(C + 1) \cdot |Z|$. Each topic is defined as a multinomial distribution over a fixed vocabulary. Particularly, the collection-specific topics $\boldsymbol{\phi}^S$ are drawn from a collection-specific Dirichlet distribution $Dir(\boldsymbol{\beta}^S)$ while the common topics

$\boldsymbol{\phi}^C$ are drawn from a collection-independent Dirichlet distribution $Dir(\boldsymbol{\beta}^C)$. Each collection-topic pair (c, z) is associated with a Bernoulli distribution with the parameter $\boldsymbol{\psi}$. The parameter $\boldsymbol{\psi}$ follows a Beta distribution $Beta(\boldsymbol{\gamma})$, and indicates how likely a word is assigned to a common topic. The hyper-parameter $\boldsymbol{\gamma}$ is the prior knowledge of $\boldsymbol{\psi}$. For each word d_i in document d, we draw a switching variable $x_{d_i} \sim Bern(\boldsymbol{\psi})$ which is a binary random variable for deciding whether a topic is collection-independent or collection-specific. Similar with the LDA, each document d has a topic proportion $\boldsymbol{\theta}_d \sim Dir(\boldsymbol{\alpha}_c)$ over the shared topic indices. Different from the CCLDA, our model embeds the logistic regression model for incorporating the class labels accompanied with documents in the training source domain. In particular, each observed class label $y_d \in \{-1, 1\}$ is drawn from $Bern(logistic(-y_d \boldsymbol{\eta}^T \overline{\boldsymbol{z}}_d))$ where $logistic(t) = \dfrac{1}{1 + e^{-t}}$ is a logistic function.

3.3.2 Learning Algorithm

Exact inference of topic models is often intractable. To learn our model, we employ a stochastic EM framework (Doyle and Elkan, 2009; Hong et al., 2011) which combines the functional optimization problem with Gibbs sampling. Specifically, in E-steps, we fix the logistic coefficients $\boldsymbol{\eta}$, and sample the hidden variables z and x, gather useful counts and update the new document representation $\overline{\boldsymbol{z}}_d$. In M-steps, we update the logistic regression coefficients $\boldsymbol{\eta}$ by maximizing the joint likelihood of all observed data and hidden variables, which is equivalent to minimizing the objective function of the associated logistic regression model. In this way, our model follows the theory in (Ben-david et al., 2006) to induce the new topic feature representation by explicitly minimizing the difference between the source and target domains and the empirical training errors in the source domain.

In the following, we elaborate the update formulas in E-step and M-step of our learning algorithm.

E-step

Let hyper-parameters $\{\boldsymbol{\alpha}_c, \boldsymbol{\beta}^C, \boldsymbol{\beta}^S, \boldsymbol{\gamma}\}$ be denoted as $\boldsymbol{\Psi}$, and hidden variables $\{\boldsymbol{\phi}^C, \boldsymbol{\phi}^S, \boldsymbol{\theta}, \boldsymbol{\psi}\}$ as $\boldsymbol{\Phi}$. The joint distribution of the observed variables and hidden variables, after integrating out $\boldsymbol{\Phi}$, is:

$$\begin{aligned}
&p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{c}, \boldsymbol{y} | \boldsymbol{\Psi}, \boldsymbol{\eta}) \\
=&\int p\left(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{c}, \boldsymbol{y}, \boldsymbol{\Phi} | \boldsymbol{\Psi}, \boldsymbol{\eta}\right) d\boldsymbol{\Phi} \\
=&\prod_c \prod_z \frac{B\left(\boldsymbol{\gamma} + \boldsymbol{n}^x_{c,z}\right)}{B\left(\boldsymbol{\gamma}\right)} \prod_d \frac{B\left(\boldsymbol{\alpha}_{c_d} + \boldsymbol{n}^z_d\right)}{B\left(\boldsymbol{\alpha}_{c_d}\right)} \\
&\prod_z \frac{B\left(\boldsymbol{\beta}^C + \boldsymbol{n}^w_{z,x=0}\right)}{B\left(\boldsymbol{\beta}^C\right)} \prod_c \prod_z \frac{B\left(\boldsymbol{\beta}^S + \boldsymbol{n}^w_{c,z,x=1}\right)}{B\left(\boldsymbol{\beta}^S\right)} \\
&\prod_d \frac{1}{1 + e^{-y_d \cdot \boldsymbol{\eta}^T \overline{\boldsymbol{z}}_d}}
\end{aligned} \tag{3.1}$$

Here, $B(\boldsymbol{\cdot})$ denotes the function $B(\boldsymbol{v}) = \dfrac{\prod_i \Gamma(\boldsymbol{v}_i)}{\Gamma(\sum_i \boldsymbol{v}_i)}$, where $\boldsymbol{v}$ is a vector and $\Gamma(\cdot)$ is the gamma function. $\boldsymbol{n}^x_{c,z}$ is a 2-dimensional vector $(n^{x=0}_{c,z}, n^{x=1}_{c,z})$ whose elements are the number of word tokens d_i in collection c which satisfy $z_{d_i} = z$ & $x_{d_i} = 0$ and $z_{d_i} = z$ & $x_{d_i} = 1$ respectively. Similarly, $\boldsymbol{n}^z_d$ is a $|Z|$-dimensional vector where each element is the number of word tokens d_i in document d which satisfies $z_{d_i} = z$. $\boldsymbol{n}^w_{z,x=0}$ is a $|W|$-dimensional vector where each element is the number of word tokens w assigned to the common topic z. $\boldsymbol{n}^w_{c,z,x=1}$ is a $|W|$-dimensional vector where each element is the number of word tokens w assigned to the collection-specific topic z in the collection c.

In E-steps, we conduct the collapsed Gibbs sampling using the following updating formulas which are derived based on equation 3.1:

$$\begin{aligned}
&p(z_i|\boldsymbol{z}_{-i}, \boldsymbol{x}, \boldsymbol{w}, \boldsymbol{c}, \boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\eta}) \\
&\propto \frac{1 + e^{-y_d \cdot \boldsymbol{\eta}^T \overline{\boldsymbol{z}}_d} e^{y_d \cdot \eta_{z_i}/n_d}}{1 + e^{-y_d \cdot \boldsymbol{\eta}^T \overline{\boldsymbol{z}}_d}} \times \left(n_d^{z_i,-i} + \alpha_{c_d z_i} \right) \\
&\times \begin{cases} \dfrac{n_{z_i,x_i}^{w_i,-i} + \beta^C}{n_{z_i,x_i}^{(\cdot),-i} + W\beta^C}, & \text{if } x_i = 0 \\ \dfrac{n_{c_d,z_i,x_i}^{w_i,-i} + \beta^S}{n_{c_d,z_i,x_i}^{(\cdot),-i} + W\beta^S}, & \text{if } x_i = 1 \end{cases}
\end{aligned} \tag{3.2}$$

$$\begin{aligned}
&p(x_i|\boldsymbol{x}_{-i}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{c}, \boldsymbol{y}, \boldsymbol{\Psi}, \boldsymbol{\eta}) \\
&\propto \frac{1 + e^{-y_d \cdot \boldsymbol{\eta}^T \overline{\boldsymbol{z}}_d} e^{y_d \cdot \eta_{z_i}/n_d}}{1 + e^{-y_d \cdot \boldsymbol{\eta}^T \overline{\boldsymbol{z}}_d}} \times \left(\frac{n_{c_d,z_i}^{x_i,-i} + \gamma_{x_i}}{n_{c_d,z_i}^{(\cdot),-i} + \gamma_{x_0} + \gamma_{x_1}} \right) \\
&\times \begin{cases} \dfrac{n_{z_i,x_i}^{w_i,-i} + \beta^C}{n_{z_i,x_i}^{(\cdot),-i} + W\beta^C}, & \text{if } x_i = 0 \\ \dfrac{n_{c,z_i,x_i}^{w_i,-i} + \beta^S}{n_{c,z_i,x_i}^{(\cdot),-i} + W\beta^S}, & \text{if } x_i = 1 \end{cases}
\end{aligned} \tag{3.3}$$

Here, the superscript $-i$ denotes a counting variable which excludes the i-th word index in the corpus, and the superscript $(\cdot)$ denotes a counting variable which sums over all elements in the corresponding vector. $n_d^{z_i,-i}$ is the number of word tokens assigned to the topic z_i in the document d, excluding the current word index i. $n_{c_d,z_i}^{x_i,-i}$ is the number of word tokens assigned to the collection-independent (if $x_i = 0$) or collection-specific (if $x_i = 1$) topic z_i in the collection c_d, excluding the current word index i. $n_{z_i,x_i}^{w_i,-i}$ and $n_{c,z_i,x_i}^{w_i,-i}$ is the number of times that the word token w_i is assigned to the collection-independent and the collection-specific topic z_i in the collection c respectively. We assume that the elements of the vector $\boldsymbol{\beta}^C$ and $\boldsymbol{\beta}^S$ have the identical value β^C and β^S respectively.

M-step

In M-steps, we update the logistic coefficients $\boldsymbol{\eta}$ by maximizing the joint likelihood in Equation 3.1. Since we fix the counts gathered in the E-step, this is equivalent to learning a new logistic regression model where each document d is represented by $\overline{z}_d$ newly updated in the E-step. Specifically, we learn a L2-regularized logistic regression model which solves the following unconstrained optimization problem:

$$\min_{\boldsymbol{\eta}} \frac{1}{2}\boldsymbol{\eta}^T\boldsymbol{\eta} + R\sum_{d} \log\left(1 + e^{-y_d \cdot \boldsymbol{\eta}^T \overline{z}_d}\right) \tag{3.4}$$

where R is a regularization parameter and set to 1.0 in our model. We apply the trust region Newton method (Lin et al., 2008; Fan et al., 2008) for optimization.

Overall Algorithm for Cross-Domain Learning

Having described the update formulas in E-steps and M-steps, we now present the algorithm for leveraging our model for cross-domain text classification.

Algorithm 1 PSCCLDA for Cross-Domain Text Classification

```
Input: labeled training data in the source domain $\mathcal{D}^s$; unlabeled testing data in the
    target domain $\mathcal{D}^t$; number of topics $K$; number of iterations $T_{em}$ and $T_{gibbs}$
Output: predicted class label of each unlabeled document $d$ in target domain $\mathcal{D}^t$
Initialize hidden variables $\boldsymbol{z}$ and $\boldsymbol{x}$ in the model
for $t := 1 \rightarrow T_{em}$ do
    E-step:
    for $t_e := 1 \rightarrow T_{gibbs}$ do
        Run collapsed Gibbs sampling for all documents using Equation (3.2) and (3.3)
    end for
    Update topic frequency $\overline{\boldsymbol{z}}_d$ for each document $d$
    M-step:
    Update logistic regression coefficients $\boldsymbol{\eta}$ using Equation (3.4)
end for
Predict class label of each document $d$ in target domain $\mathcal{D}^t$ using Equation (3.5)
```

The overall procedure is depicted in Algorithm 1. We first learn our model by alternately running E-step and M-step for T_{em} iterations (or until convergence). Each domain is treated as a collection, and the model is initialized randomly. In E-steps, we run Gibbs sampling for T_{gibbs} iterations. During the sampling, if a document is labeled (i.e., from the training source domain), we use the exact update formulas in Equation 3.2 and 3.3; if a document is unlabeled, (i.e., from the target domain), we use the update formulas in Equation 3.2 and 3.3 by removing the first term containing the class label y_d. When the sampling is finished, we update the empirical topic frequency $\overline{\boldsymbol{z}}_d$ using the last sample obtained. In particular, $\overline{\boldsymbol{z}}_d$ is a $2 * |Z|$ dimensional vector where the first $|Z|$ dimensions correspond to the collection-independent topics $z^C \in \{1...Z\}$ in turn, and the remaining $|Z|$ dimensions correspond to the collection-specific topics $z^S \in \{1...Z\}$ in turn. Each element in $\overline{\boldsymbol{z}}_d$ is the normalized frequency of the corresponding topic n_z/n_d where n_d is the total number of tokens in the document d. In M-steps, we update the logistic coefficients $\boldsymbol{\eta}$ using the Equation 3.4.

When the model is learned, we can directly use the last updated $\overline{\boldsymbol{z}}_d$ of unlabeled documents and logistic coefficients $\boldsymbol{\eta}$ for predicting. Specifically, we predict the class label of the unlabeled document d using the following equation:

$$p(y_d|\overline{\boldsymbol{z}}_d) = \frac{1}{1 + e^{-y_d \cdot \boldsymbol{\eta}^T \overline{\boldsymbol{z}}_d}} \tag{3.5}$$

If $p(y_d = 1|\overline{\boldsymbol{z}}_d) \geq 0.5$, the predicted class label is 1; otherwise, the predicted class label is -1.

3.4 Experiments

In this section, we evaluate the effectiveness of our proposed PSCCLDA model for the problem of cross-domain text classification in general, and cross-industry sentiment analysis in particular.

3.4.1 Data Preparation

In the following, we describe the data preparation for the evaluation on two tasks, including the cross-domain text classification and the cross-industry risk sentiment analysis.

Data Preparation for Cross-Domain Text Classification

To evaluate the performance for cross-domain text classification, we use the nine datasets provided by Li et al. (2012) [2], which are generated from two widely used text classification datasets, i.e., 20Newsgroups [3] and Reuters-21578 [4]. The nine datasets are generated by using the hierarchical category structures in the same way as many previous studies on cross-domain learning (Li et al., 2012; Xue et al., 2008; Zhuang et al., 2010; Pan and Yang, 2010; Jiang, 2008; Long et al., 2012). Specifically, both the 20Newsgroups and the Reuters-21578 datasets are organized under top categories, e.g., *comp* (computer), *rec* (recreation), *sci* (science), and *talk* in the 20Newsgroups dataset, and *orgs*, *people*, *places* in the Reuters-21578 dataset. Each top category contains sub-categories. For example, under the top category "comp", there are sub-categories such as "comp.graphics", "comp.os.ms-windows.misc", "comp.sys.ibm.pc.hardware" and "comp.sys.mac.hardware". The datasets for cross-domain text classification are generated as follows. Suppose we have two top-categories A and B, which contain four sub-categories A_1, A_2, A_3, A_4 and B_1, B_2, B_3, B_4 respectively. The task is defined as the top-category binary

[2] Available at http://www.cse.ust.hk/TL/index.html.

[3] Available at http://people.csail.mit.edu/jrennie/20Newsgroups

[4] Available at http://www.daviddlewis.com/resources/testcollections

classification in the cross-domain context, i.e., to classify documents into either A or B. To generate the training data in the source domain, we randomly choose two sub-categories for each top-category, and merge all documents in the chosen sub-categories (e.g., A_1, A_2, B_1, B_2). The documents in the remaining sub-categories are merged as the testing data in the target domain (e.g., A_3, A_4, B_3, B_4). In this way, the data in the source domain and target domain are related (since they come from the same top-category) but distributed differently (since they come from the different sub-categories).

Table 3.2 shows the statistics of the nine datasets generated as described above. In the "Dataset" column, we list the name of each dataset based on the defined classification task. For example, "comp vs rec" denotes the dataset in which the task is the binary classification of two top categories "comp" and "rec". In the "Instances" column, we present the number of instances for both labeled data (l) in the source domain and unlabeled data (u) in the target domain. In the "Features" column, we show the distributional difference between domains where $|F_{s-t}|$ is the number of word features that exclusively appeared in the source domain, and $|F_{s\cap t}|$ is the number of word features that appears in both the source and target domains. A larger ratio $|F_{s-t}|/|F_{s\cap t}|$ indicates more salient distributional difference between domains.

Data Preparation for Cross-Industry Risk Sentiment Analysis

To prepare the data for the evaluation on cross-industry risk sentiment analysis, we collect and extract the "Management Discussion and Analysis" (MD&A) section in 10-K forms. The MD&A is one of the most examined sections in the 10-K form, and is mandated by the Securities and Exchange Commission (SEC) since 1980. This disclosure section is intended to access a company's finical condition

Table 3.2: Statistics of datasets for cross-domain text classification.

Dataset	Instances		Features		
	l	u	$\lvert F_{s-t}\rvert$	$\lvert F_{s\cap t}\rvert$	$\lvert F_{s-t}\rvert/\lvert F_{s\cap t}\rvert$
comp vs rec	3933	3904	3475	11646	0.2984
comp vs sci	3911	3901	3092	12564	0.2461
comp vs talk	3654	3464	4434	13654	0.3247
rec vs sci	3591	3958	4456	13194	0.3378
rec vs talk	3690	3525	4954	13867	0.3573
sci vs talk	3373	3818	3698	15129	0.2444
orgs vs people	1237	1208	230	4090	0.0562
orgs vs places	1061	1043	178	3892	0.0457
people vs places	1077	1077	233	3833	0.0608

and results of operations, and allows company management to tell its story in its own words in a way that investors can understand. According to the SEC (2014), the MD&A section should present "the company's operations and financial results, including information about the company's liquidity and capital resources and any known trends or uncertainties that could materially affect the company's results", and may also discuss "management's views of key business risks and what it is doing to address them". In our experiment, the task is to predict the risk sentiment embodied in the MD&A section in the context of cross-industry learning.

Specifically, we collect the 10-K forms from 1996 to 2006, and then extract the MD&A section in 10-K forms as disclosure documents. Each document is processed by lowercasing and removing punctuations, stopwords, and meaningless marks. The membership of industry is determined by SIC (Standard Industrial Classification) code as in (Frankel et al., 2002). Since we have no ground-truth labels of risk sentiment, we use the stock return volatility (SRV) as a proxy for risk (Kogan et al., 2009) – higher SRV indicates higher level of risk. We assume that the MD&A section is informative for investors (Li, 2010a), and the market will react to it after the filing. Therefore, we label a document as "positive" if the SRV 12 months before the filing date (SRV^{-12}) is smaller than the SRV 12 months after the filing date (SRV^{+12}), and "negative" otherwise.

Figure 3.3 shows the distribution of the number of observations (i.e., documents) across industries. To reduce the imbalance of the dataset, we only retain 7 out of 13 industries whose number of observations are larger than 1600, including "computer", "extractive", "manufacture", "pharmaceutical", "retail", "service", and "transportation" industry.

We create 42 tasks of the cross-industry risk sentiment analysis by paring the 7 industries in our sample. For each industry pair, one industry is regarded as the source industry (for training), and the other one is regarded as the target industry (for testing).

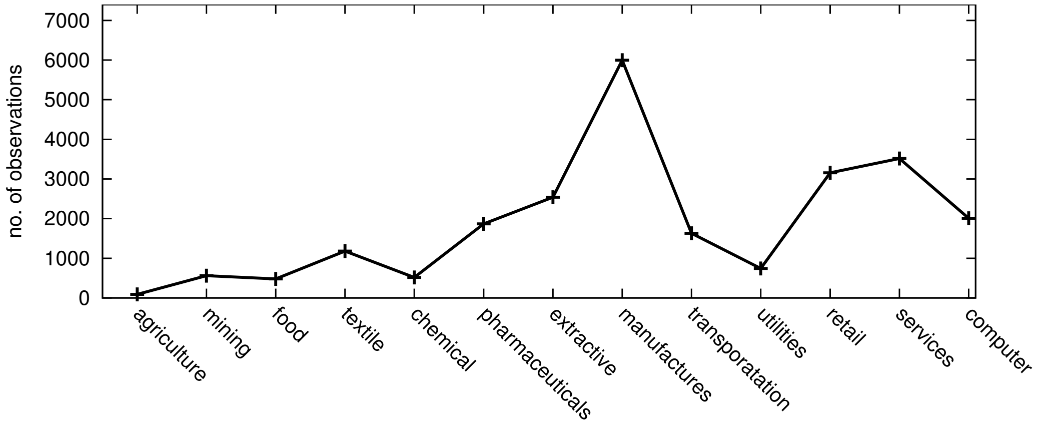


Figure 3.3: Data distribution across industries.

3.4.2 Experimental Settings

We now describe the experimental settings, including the benchmark methods and the performance metric.

The benchmark methods used in our experiments include two conventional supervised classification algorithms, namely the SVM (Support Vector Machine) implemented in (Fan et al., 2008) and the LG (Logistic Regression) implemented in (Fan et al., 2008), and four state-of-the-art cross-domain text classification algorithms that have been reviewed in Section 2.3.4, including the SFA (Spectral Feature Alignment) (Pan et al., 2010), the TPLSA (Topic-bridge PLSA) (Xue et al., 2008), the CDPLSA (Collaborative Dual-PLSA) (Zhuang et al., 2010) and the TCA (Topic Correlation Analysis) (Li et al., 2012).

In our evaluation on the cross-domain text classification, we aim to demonstrate the superiority of our proposed model over existing cross-domain learning methods. Therefore, we compare our model with both the two conventional supervised classification algorithms and the four state-of-the-art cross-domain text classification algorithms. On the other hand, in our evaluation on the cross-industry risk sentiment analysis, we aim to show that our model could outperform the existing methods that have been adopted in the financial accounting domain. Therefore, we only compare our model with the two conventional supervised learning methods.

For the conventional supervised algorithms SVM and LG, we perform the classification in a traditional way. Specifically, we train the classifier using the labeled documents in the source domain and directly use the trained model to predict the class labels of unlabeled documents in the target domain. The parameters are set to the default values as in Fan et al. (2008). The performance of these two conventional classifiers serve as the baselines for the cross-domain learning methods. For the competing cross-domain learning methods, we mainly choose the topic modeling based methods because we focus on exploring how to extend original topic models for cross-domain learning in this study. The parameters are set as in the original papers (Pan et al., 2010; Xue et al., 2008; Zhuang et al., 2010; Li et al., 2012).

To measure the performance, we use the common metric *classification accuracy*, which is defined as the proportion of correctly classified examples (i.e., documents).

3.4.3 Evaluation on Cross-Domain Text Classification

We now present the experimental results of the evaluation on the cross-domain text classification.

Overall Performance

In Table 3.3, we show the performance comparison of our proposed PSCCLDA model with all benchmark methods, including the two conventional supervised classifiers and the four competing cross-domain learning methods, on all the nine datasets. Since our model is randomly initialized, we run the model 3 times and report the "mean ± standard deviation" in the column "PSCCLDA". For our model, the parameters are tuned on the "comp vs rec" dataset and then applied to all the other datasets. In particular, the number of the topic indices $|Z|$ is set to 5 and 6 for the datasets generated from 20Newsgroups and Reuters-21578 respectively; γ is set to $(20, 1)$, indicating that the domain-independent topics ($x = 0$) are more likely to be chosen than the collection-specific topics ($x = 1$); the number of EM iterations T_{em} is set to 50 and the number of iterations for Gibbs sampling T_{gibbs} is set to 6.

As can be seen in Table 3.3, our PSCCLDA model performs best on 7 out of 9 datasets. Two exceptions are that the SFA performs best on the "comp vs talk" dataset and the TCA performs best on the "rec vs talk" dataset. Not surprisingly, all the cross-domain learning methods perform better than the conventional supervised classifiers (i.e., LG and SVM), demonstrating that the distributional difference between domains indeed deteriorates the performance of the supervised classifiers. On average, our model outperforms all the other methods with the classification accuracy 88.0%. We also conduct the *t*-test at the 95% confidence level over all nine datasets, and the tests show that the performance improvement of our model over the benchmark methods is statistically significant (p-value < 0.05).

Table 3.3: Performance comparison for cross-domain text classification.

Datasets	LG	SVM	SFA	TPLSA	CDPLSA	TCA	PSCCLDA
comp vs rec	0.906	0.895	0.939	0.910	0.914	0.940	**0.958**±0.012
comp vs sci	0.759	0.719	0.830	0.802	0.877	0.891	**0.900**±0.014
comp vs talk	0.911	0.898	**0.971**	0.938	0.955	0.967	0.967±0.005
rec vs sci	0.719	0.696	0.885	0.928	0.872	0.879	**0.955**±0.016
rec vs talk	0.848	0.827	0.935	0.849	0.912	**0.962**	0.958±0.019
sci vs talk	0.780	0.747	0.854	0.890	0.862	0.940	**0.947**±0.013
orgs vs people	0.681	0.670	0.671	0.746	0.808	0.792	**0.807**±0.013
orgs vs places	0.692	0.669	0.683	0.719	0.714	0.730	**0.742**±0.036
people vs places	0.513	0.520	0.506	0.623	0.548	0.626	**0.690**±0.057
average	0.757	0.738	0.808	0.823	0.829	0.859	**0.880**

Notes: The performance is measured by the classification accuracy. The best performance is highlighted in bold font.

Convergence and Parameter Sensitivity

We now examine the convergence and parameter sensitivity of our model.

First, we need to ensure the performance convergence for our model since we use an EM-style learning algorithm. Figure 3.4 presents the model performance in

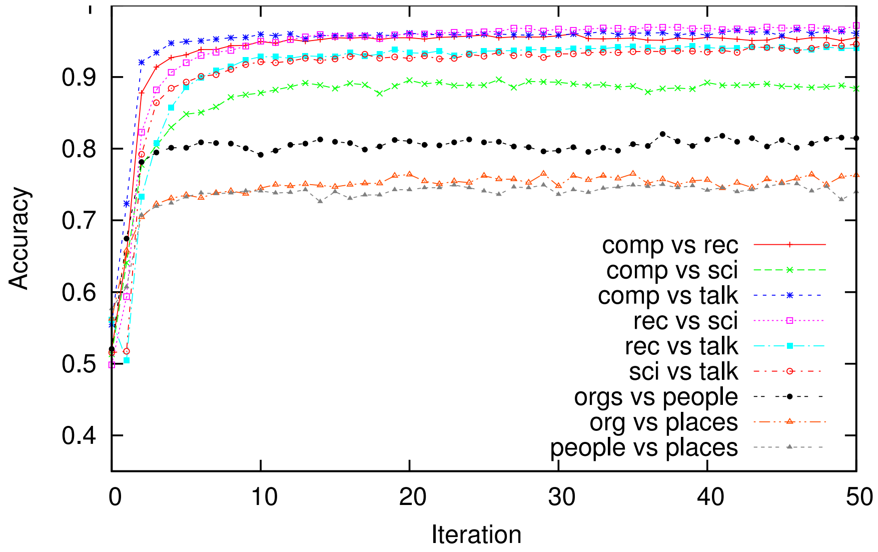


Figure 3.4: Performance convergence.

terms of the classification accuracy by varying the number of EM iterations. As can be seen, the performances of our model on all datasets increase quickly during the first 10 iterations, and then tend to converge to the constant values. This observation demonstrates that our model could ensure the convergence on all the nine datasets.

Second, we need to ensure that our model is not sensitive to the model parameters that have to be empirically set. There are two important parameters in our PSCCLDA model, including the number of topic indices $|Z|$, and the hyper-parameter γ which could be interpreted as the prior belief on the proportion of domain-independent and domain-specific topics. To investigate the effects of these two parameters, we show the performance of our model on all the datasets by varying one parameter while fixing the other one.

In Figure 3.5, we fix the parameter γ to its default value $(20, 1)$, and vary the number of topics from 2 to 20. As can be seen, the model performance is relatively stable when the number of topics is larger than 5. This demonstrates that our model is not very sensitive to the parameter $|Z|$.

In Figure 3.6, we fix the parameter $|Z|$ to 5 and 6 for datasets generated from the 20Newsgroups and Reuters-21578 datasets respectively, and vary $\gamma_{x=0}$ from 0.5 to 100 while fixing $\gamma_{x=1}$ to 1. As can be seen, the model performance is relatively stable when $\gamma_{x=0}$ is larger than 20. This demonstrates that our model is not sensitive to the parameter γ. It is interesting to notice that the performance of our model will drop when $\gamma_{x=0} \leq \gamma_{x=1}$, indicating that the collection-independent topics have more predictive power than the collection-specific topics in the context of cross-domain learning.

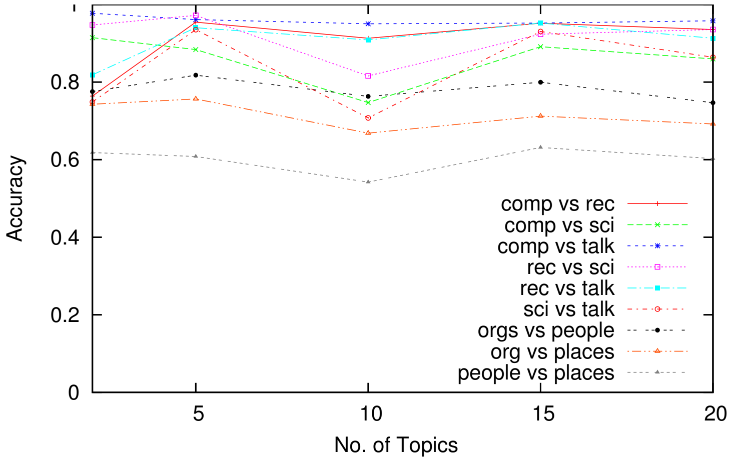


Figure 3.5: Performance of our model by varying the number of topics.

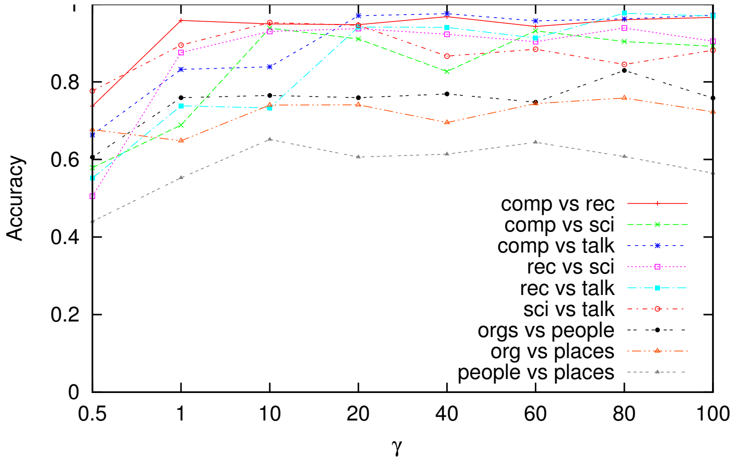


Figure 3.6: Performance of our model by varying the parameter γ.

Effects of Domain Difference

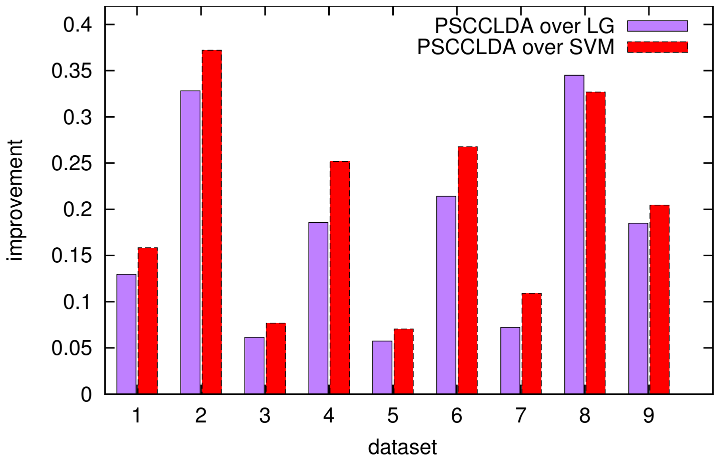


Figure 3.7: Performance improvement of PSCCLDA.

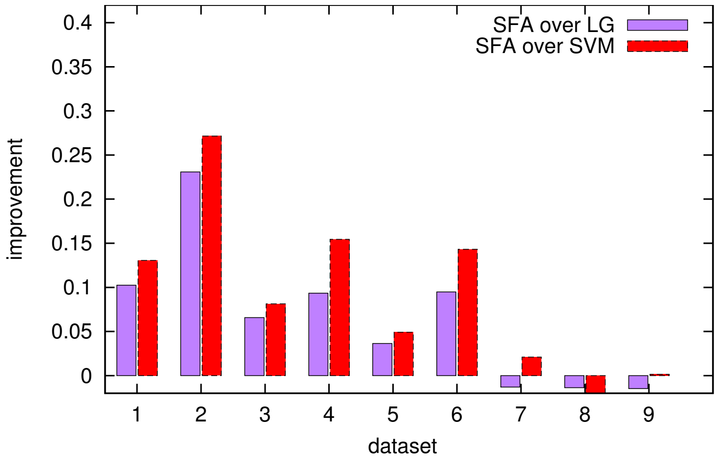


Figure 3.8: Performance improvement of SFA.

Here, we conduct an analysis on the effects of the distributional difference between domains on the relative improvement of the cross-domain learning methods over the supervised ones. We use the KL (Kullback Leibler) divergence to quantify the domain difference for the nine datasets in Table 3.2, and rank them from 1 (smallest difference) to 9 (largest difference) in an ascending order. Figure 3.7 shows the performance improvement of our PSCCLDA model over the conventional supervised methods in terms of the classification accuracy. We expect a decreasing improvement from the dataset 1 to 9, since it is intuitive that a salient difference will indicate a difficult cross-domain learning task. However, our observation is contradictory to this intuition, which implies that the relative improvement of the cross-domain learning methods over the supervised ones may not be dependent on the distributional difference between domains. To further verify this observation, we examine the performance improvement of the SFA model, which is another type of cross-domain learning method as reviewed in Section 2.3.4. Specifically , as shown in Figure 3.8, the performance improvements of the SFA model over the supervised classifiers are not dependent on the distributional difference between domains. This observation is consistent with prior works (Dai et al., 2007), but has no theoretical explanation yet. We believe that more deep analysis on this counter-intuitive observation is needed in future, and the explanation for it will shed light on the design of more robust cross-domain learning methods.

3.4.4 Evaluation on Cross-Industry Risk Sentiment Analysis

Here, we present the experimental results of the evaluation on cross-industry risk sentiment analysis. Figure 3.9 shows the performance comparisons between our PSCCLDA model and two conventional supervised learning methods (i.e., SVM and LG) for 42 tasks of cross-industry risk sentiment analysis as described in Section 3.4.1. We observe that the conventional supervised learning methods perform poorly in the context of cross-domain learning – the SVM achieves 52.60% classification accuracy on average while the LG achieves 52.94% on average. This is only slightly better than the random guess (50% accuracy). In contrast, our PSCCLDA model could lead to roughly 6% performance improvement over the supervised learning methods, achieving 58.14% classification accuracy on average. This demonstrably shows the effectiveness and superiority of our proposed PSCCLDA model over the existing supervised learning methods for the cross-industry risk sentiment analysis.

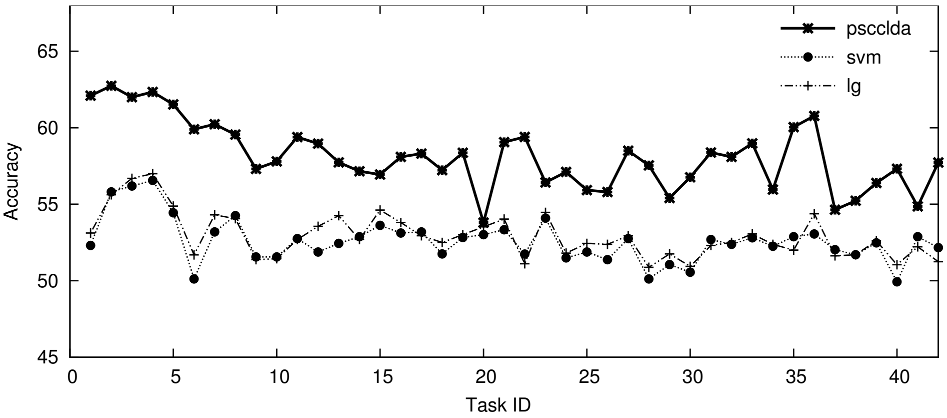


Figure 3.9: Performance comparisons for cross-industry risk sentiment analysis.

3.5 Summary

In this chapter, we study the problem of cross-domain text classification in general, and cross-industry risk sentiment analysis in particular. To solve the problem, we propose an extended LDA topic model, called PSCCLDA, and its learning algorithm. With the purpose of overcoming two observed limitations of the existing methods, our proposed model explicitly distinguishes the domain-independent and domain-specific topics by resorting to the cross-collection topic models, and exploits the label information for inferring more predictive topics by embedding the supervised logistic regression model. Experimental results on nine standard datasets demonstrate the effectiveness of our model for cross-domain text classification in general, and its superiority over the state-of-the-art cross-domain learning methods. Experiential results of the 42 tasks for cross-industry risk sentiment of the MD&A disclosures show the effectiveness of our proposed model, and its superiority over the existing supervised learning methods that have been adopted in the financial accounting domain.

Chapter 4

Extracting Individual Risk Types

In this chapter, we present an extended LDA model and its learning algorithm for extracting individual risk types from corporate risk disclosures without pre-defining them.

4.1 Overview

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 called 10-K form. In addition to the quantitative financial data detailed in these reports, one of the most analyzed elements in the 10-K form are 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 requires that all firms include a separate section (section 1A) in their 10-K form to discuss "the most significant factors that make the company 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 study, 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 from large amount 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 pre-defined 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. As reviewed in Section 2.3, this work falls into two categories of automated text analysis: *dictionary based* and *supervised learning based.*

Dictionary and supervised learning methods assume a pre-defined 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 tone 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 pre-defined risk types. The drawback of this assumption is further indicated by the salient difference between pre-defined risk types defined in (Mirakur, 2011; Campbell et al., 2014; 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 study, we report the first general work on extracting individual risk types from textual disclosures without pre-defining them. Specifically, we propose an unsupervised topic model which could estimate rather than pre-define a set of categories (risk types) and simultaneously assign sentences (risk factors) to those categories.

The rest of this chapter is organized as follows. Section 4.2 describes the problem formulation. Section 4.3 elaborates our proposed model and its learning algorithm. Section 4.4 presents the experiments for evaluating the proposed model. Finally, Section 4.6 provides a brief summary of the study in this chapter.

4.2 Problem Formulation

Given a collection of documents containing disclosed risk factors, our task is to (1) estimate a set of risk types at the collection level, and (2) simultaneously map each risk factor to the most suitable risk type. To get a feel for our problem, consider Apple Inc.'s 10-K form in 2006[1]. In section 1A of this form, the summary headings of three sample risk factors are listed in Table 4.1 [2]. 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 4.1, our method would yield the following risk types: Lawsuits (RT1), Catastrophes (RT2) and Human Resources (RT3). Further it would map the first risk factor in Table 4.1 to RT1, the second factor to RT2 and the last factor to RT3. Finally, this disclosure document (if only contains these three risk factors) could be quantified as a vector $[1, 1, 1]$ where each dimension corresponds to a risk type.

Table 4.1: Three sample risk factors in a disclosure document.

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.

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

[2] It should be noted that we focus on the analysis of the summary headings of these risk factors in section 1A "Risk factors" of 10-K forms, but ignore their detailed explanations. The unit of analysis in our problem is the sentence since each risk factor is described using only one sentence in most cases.

4.3 Proposed Model

In this section, we elaborate our proposed model and its learning algorithm.

4.3.1 Model Description

Our idea for solving the formulated problem is to make use of the topic model, hoping that the discovered topics are meaningful for representing risk types. Once such model is learned, individual risk types (i.e., topics) can be automatically discovered from the collection of disclosure documents, and each word (or sentence) can be assigned to the most probable risk types (i.e., topics) (Recall that one output of topic model is the topic assignments for words in each document). However, as we will show later in Section 4.4, the original LDA topic model is not adequate because the discovered topics are not meaningful for representing risk types. To address this issue, our strategy is to extend the LDA model by incorporating appropriate additional information with the purpose of steering the model towards topics that are meaningful for representing risk types.

We first elaborate the intuition behind our model. 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. This assumption is illustrated on the left side of Figure 4.1, where each dash denotes a word, each color denotes a topic, and connected dashes represent a sentence. Specifically, LDA model assumes that each word can belong to any topic based on the document-wide topic proportion (i.e., a multinomial distribution). This implies that it makes no difference whether or not two words are in the same sentence. But this assumption is clearly unrealistic in our case, since we observe, as will show in Section 4.4.1 later, that each sentence in a document is only regarding one risk type (i.e., 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 towards 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.

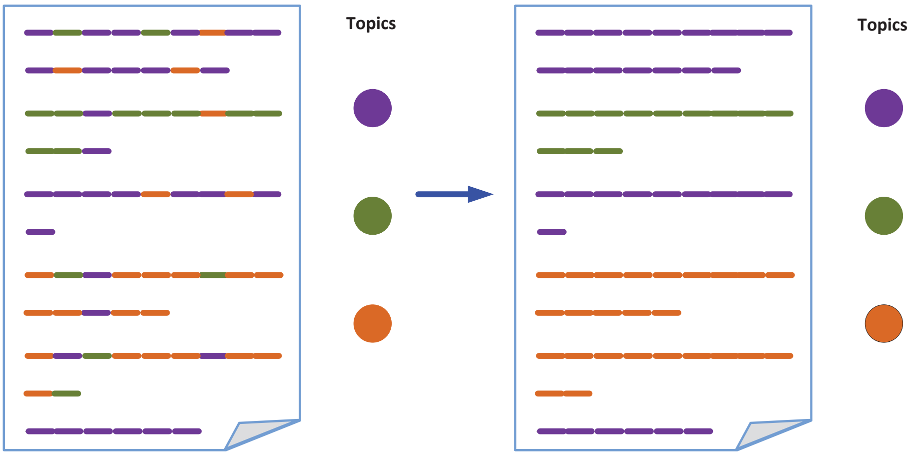


Figure 4.1: Intuition of our Sent-LDA model.

Based on our intuition, 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 "one-topic-per-sentence" assumption is illustrated on the right side of Figure 4.1. 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, as reviewed in Section 2.2, do exploit sentence structures to enhance the LDA model. Distinct from our proposed "one-topic-per-sentence" assumption, all those methods allow each sentence to include multiple topics, and use different methods to incorporate sentence structure.

We now proceed to describe our proposed model, called Sent-LDA (Sentence-based

LDA). Let M, S, N, K, V be the number of documents in a corpus, the number of sentences in a document, the number of words in a sentence, the number of topics and the vocabulary size, respectively. $Dirichlet(\cdot)$ is a Dirichlet distribution with parameter $(\cdot)$ and $Multinomial(\cdot)$ is a multinomial distribution with parameter $(\cdot)$. $\boldsymbol{\beta_k}$ is the V-dimensional word distribution for topic k, and $\boldsymbol{\theta_d}$ is the K-dimensional topic proportion for document d. $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$ are the hyper-parameters of the corresponding Dirichlet distributions. The generative process of our Sent-LDA is changed to:

1. For each topic $k \in \{1, ..., K\}$:
   (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 $\boldsymbol{z_{d,s}} \sim Multinomial(\boldsymbol{\theta_d})$
      ii. For each word $w_{d,s,n}$ in sentence s:
         A. Draw a word $w_{d,s,n} \sim Multinomial(\boldsymbol{\beta_{z_{d,s}}})$

Figure 4.2 presents the graphical representation of our Sent-LDA model, which adds a sentence layer in the original hierarchy of LDA in Figure 2.4.

4.3.2 Learning Algorithm

We now present the learning algorithm for our proposed model. As reviewed in Section 2.1.2, there are two commonly used learning algorithms, including Gibbs sampling and variational EM. There are many discussions on the advantages and

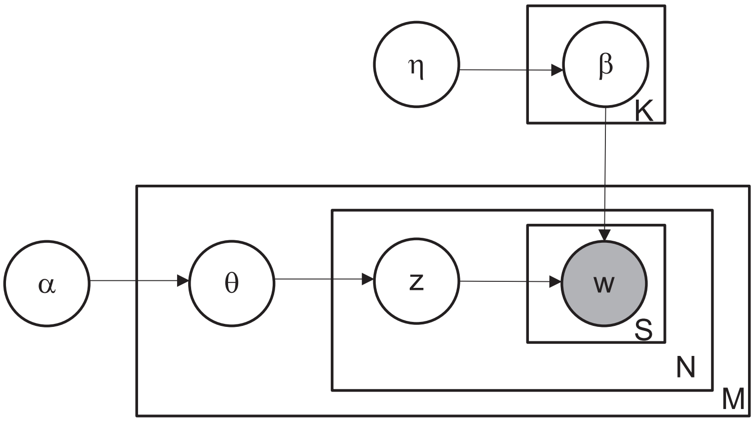


Figure 4.2: Graphical representation of Sent-LDA model.

disadvantages of them, and some previous studies (Teh et al., 2007; Asuncion et al., 2009; Wallach et al., 2009a; Zhai et al., 2012) have attempted to compare their performance. However, the findings are mixed. Following Blei and Jordan (2006), we resort 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 (CGS) method for learning. As we will demonstrate later, this Sent-LDA-CGS model performs even worse than the original LDA model for our problem. In contrast, we propose the variational EM learning algorithm for Sent-LDA model which performs best among competing methods.

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 as same as that of LDA shown in Equation 2.1. However, the interpretation of the vector $\boldsymbol{z}$ 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, $\boldsymbol{z}$ 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 (KL) divergence to the true posterior. In Sent-LDA model, the latent variables are the per-document topic proportion $\boldsymbol{\theta}$ and the per-sentence topic assignment $\boldsymbol{z}$. Similar to variational method for LDA, we use the following variational distribution:

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

as a surrogate for the posterior distribution in Equation 2.1.

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:

$$\begin{aligned}
logp(w|\alpha, \beta) &= log \int \sum_{z} p(\theta, z, w|\alpha, \beta) d\theta \\
&= log \int \sum_{z} \frac{p(\theta, z, w|\alpha, \beta) q(\theta, z)}{q(\theta, z)} d\theta \\
&\geq \int \sum_{z} q(\theta, z) logp(\theta, z, w|\alpha, \beta) d\theta - \int \sum_{z} q(\theta, z) logq(\theta, z) d\theta \\
&= E_q[logp(\theta, z, w|\alpha, \beta)] - E_q[logq(\theta, z)] \\
&= L(\gamma, \phi; \alpha, \beta)
\end{aligned}$$

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

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

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 to 1 if word w_n is the j-th 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 LDA in (Blei et al., 2003) lies in the second, third and fifth terms due to 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:

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

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

where ϕ_{si} is the probability that sentence s is generated by topic i, and γ_i is the

i-th component of posterior Dirichlet parameter.

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 j-th word is generated by topic i, all words w_{dsn} in sentence s share the same ϕ_{si}, and w_{dsn}^{j} equals to 1 if word w_{dsn} in sentence s of document d is the j-th 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.

4.4 Experiments

In this section, we evaluate the effectiveness of our proposed Sent-LDA model for extracting individual risk types without pre-defining them.

4.4.1 Data Preparation

We first describe the data preparation for our experiments, including the data collection, training set construction, accuracy of data extraction, and validation of our key "one-topic-per-sentence" assumption.

Data Collection

To collect our dataset, 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 4.1. Due to 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 the needed information using pre-defined 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., mis-extracted content. As we will report later in this section, we

[3] http://www.sec.gov/edgar/searchedgar/ftpusers.htm

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 dataset 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.

Training Set Construction

In order to compare our unsupervised method with supervised methods, we have to construct a training set for learning supervised models. The construction of training set consists of two steps. First, we have to pre-define 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 find 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 mis-extracted 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 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.

Accuracy of Data Extraction. Due to the heuristic nature of our extraction procedure as described previously, we end up with some mis-extracted sentences of disclosures. During the manual labeling procedure, these sentences are awarded the label of "Not a Risk Type". At the end of the labeling task, we counted the number of such mis-extracted sentences, and found that only 17 of the 1842 extracted sentences (0.92%) possessed a "Not a Risk Type" label, indicating the robustness of our extraction heuristics.

Validation of "One-Topic-Per-Sentence" Assumtion. To validate our fundamental assumption that each 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% (11/1327) 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 "One-Topic-Per-Sentence" assumption of our proposed model.

4.4.2 Experimental Settings

We now describe the experimental settings for our evaluation, including benchmark methods, and their parameter settings.

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, and is demonstrated to be competitive as reviewed in Section 2.2.3. To examine the effect of learning algorithms, we learn each model with two learning algorithms, 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 while the others are benchmarks. It is worth noting that Sent-LDA model assigns topics at the sentence level while 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 1000, and the likelihood convergence criteria is 1×10^{-5}. For the collapsed Gibbs sampling, we set the hyper-parameters 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 2000.

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 the each risk factor is only regarding one risk type, we simply classify each risk factor with the most probable risk type generated by CKNN algorithm.

4.4.3 Evaluation of Model Fit

In this section, we evaluate our proposed model in terms of several objective measures of model fit that are commonly used for unsupervised topic models, including perplexity, empirical likelihood, and silhouette coefficient.

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 documents 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:

$$perplexity(D_{test}) = exp(-\sum_{d=1}^{M} logp(w_d)/\sum_{d=1}^{M} N_d)$$

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

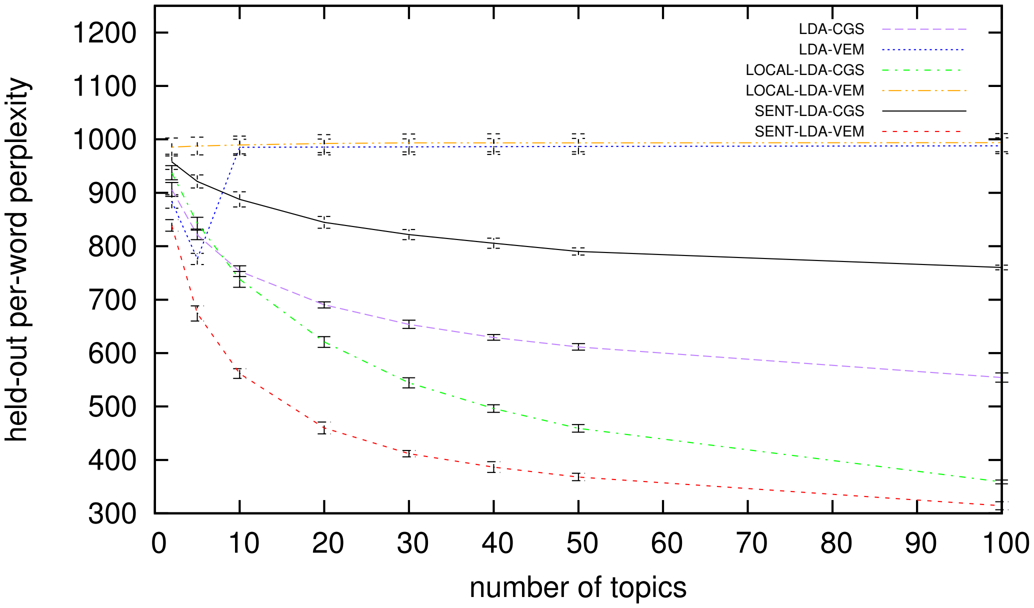


Figure 4.3: Held-out perplexity as a function of the number of topics.

Figure 4.3 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 10-fold cross validation as in (Blei and Lafferty, 2007). Specifically, we first divide the data into ten 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 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 perplexities of all methods monotonically decrease with the increase of the number of topics, but tend to converge to a fixed value eventually. When the number of topics is larger than 30, the perplexity tends to be steady.

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 4.2. As can be seen, our proposed Sent-LDA-VEM significantly outperforms all the benchmark methods at a 1% significance level.

Table 4.2: Comparison between models in terms of the held-out perplexity.

	LDA-CGS	**LDA-VEM**	**LLDA-CGS**
Mean	632.91	967.11	524.47
Std	(± 8.26)	(± 16.84)	(± 8.74)
t-value	-57.90	-92.24	-31.44
p-value	0.0000	0.0000	0.0000
	LLDA-VEM	**SLDA-CGS**	**SLDA-VEM**
Mean	973.78	804.33	**389.04**
Std	(± 18.13)	(± 10.17)	(± 10.45)
t-value	-88.39	-90.08	-
p-value	0.0000	0.0000	-

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 4.4. We also plot the model performance as a function of running time in Figure 4.6 since VEM algorithms usually need dozens of iterations to converge, while CGS algorithms require thousands of shorter iterations (Zhai et al., 2012). As shown in Figure

4.4 and 4.6, 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).

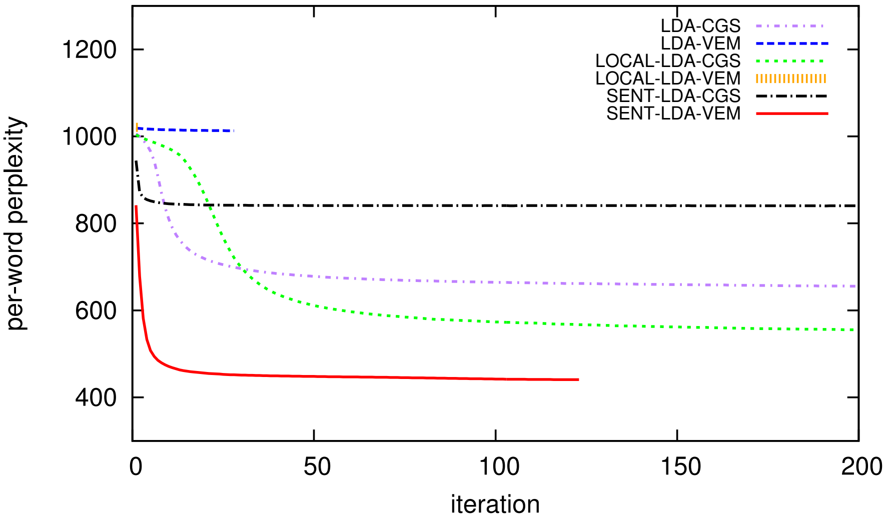


Figure 4.4: Perplexity as a function of number of iterations.

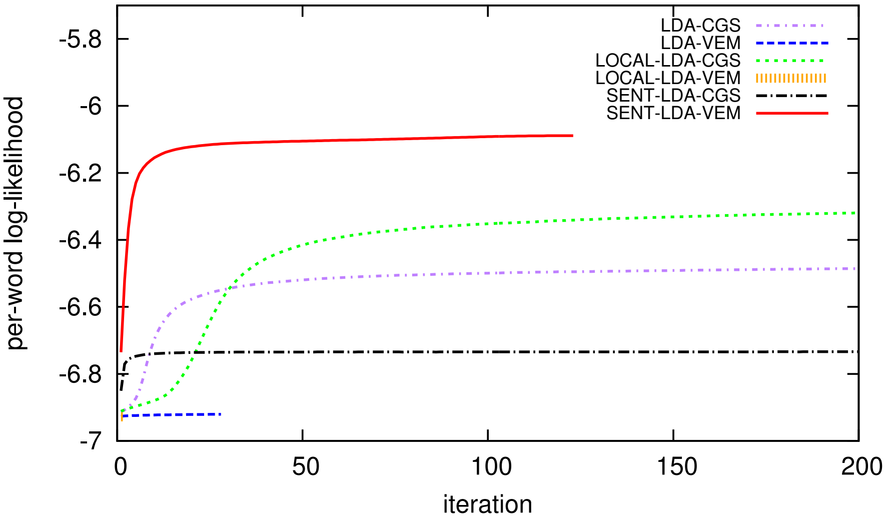


Figure 4.5: Empirical log-likelihood as a function of the number of iterations.

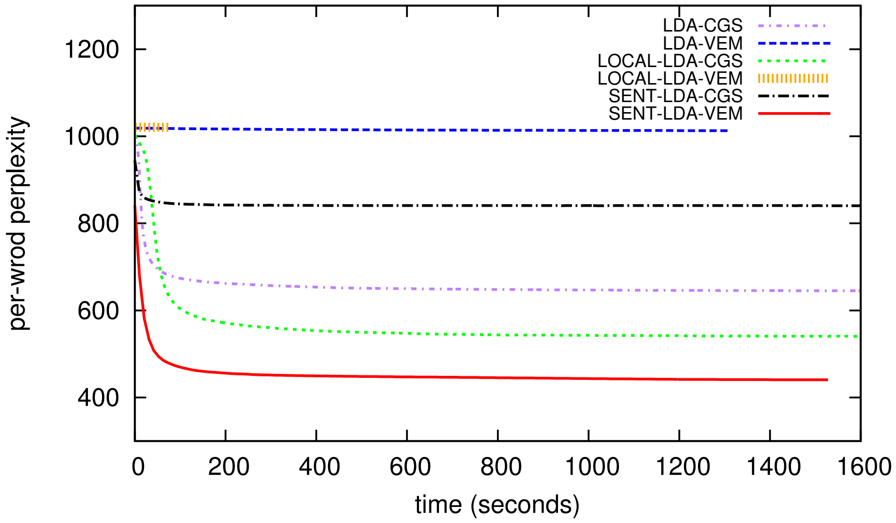


Figure 4.6: Perplexity as a function of time.

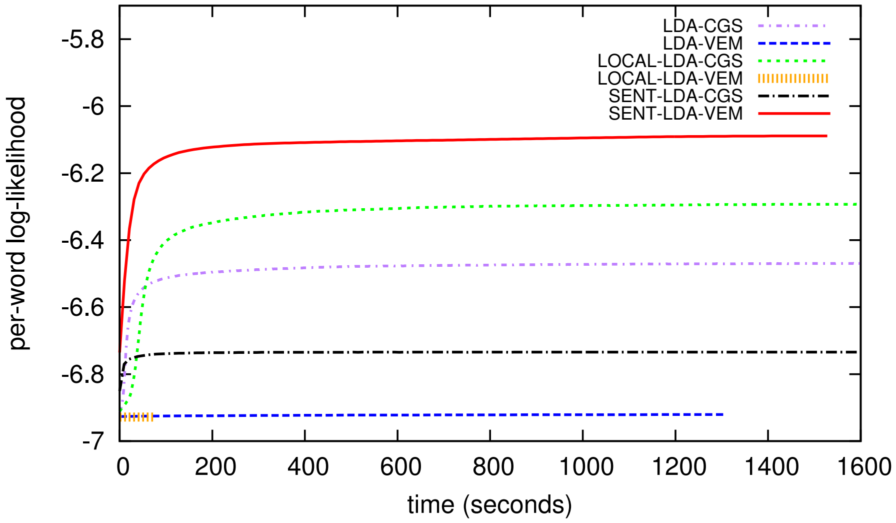


Figure 4.7: Empirical log-likelihood as a function of time.

Cluster Quality

Cluster quality refers to the extent to which intra-cluster similarities outdistance inter-cluster 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 silhouette of all points. The silhouette is bounded between -1 (for bad clustering) and $+1$ (for highly dense clustering).

Table 4.3 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 10-fold cross validation in the same way as that in Table 4.2. As can be seen, our Sent-LDA-VEM performs best among all models. It significantly outperforms LDA-CGS, LDA-VEM, Local-LDA-VEM, but performs equally well as Local-LDA-CGS and Sent-LDA-CGS.

It should be noted that silhouette coefficient assumes hard clustering while 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.

Table 4.3: Comparison between models in terms of the silhouette coefficient.

	LDA-CGS	**LDA-VEM**	**LLDA-CGS**
Mean	-0.06358	-0.08186	-0.03284
Std	(± 0.01066)	(± 0.01847)	(± 0.00786)
t-value	7.19	7.70	0.84
p-value	0.0000	0.0000	0.4112
	LLDA-VEM	**SLDA-CGS**	**SLDA-VEM**
Mean	-0.12421	-0.02975	**-0.02932**
Std	(± 0.05267)	(± 0.00673)	(± 0.01064)
t-value	5.588	0.12	-
p-value	0.0000	0.9152	-

4.4.4 Evaluation of Discovered Information

The objective measures reported in the previous section are essential for evaluating the model, and have been 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 (Chang et al., 2009; Grimmer and Stewart, 2013). To this end, we evaluate the quality of the discovered information below.

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 they are not suitable in cases where the labeling requires domain knowledge (financial knowledge in our case). Actually, in most topic model research, it is customary to manually label topics to ensure the high labeling quality (Chang et al., 2009). We thus design a manual labeling procedure which 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 the 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 4.4 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. After the independent mapping, the annotators get together to achieve consensus, and then decide the labels for topics marked with "Other risk types". Figure 4.8 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.

Here, we take Figure 4.8 as an example for illustrating how our unsupervised topic model can be used for suggesting a classification scheme for supervised

Table 4.4: Inter-rater agreement for labeling topics.

	Cohen's Kappa	Max Cohen's Kappa	p-value
LDA-CGS	0.4958	0.6639	0.0000
Local-LDA-CGS	0.6296	0.7407	0.0000
Sent-LDA-VEM	0.8400	0.8400	0.0000

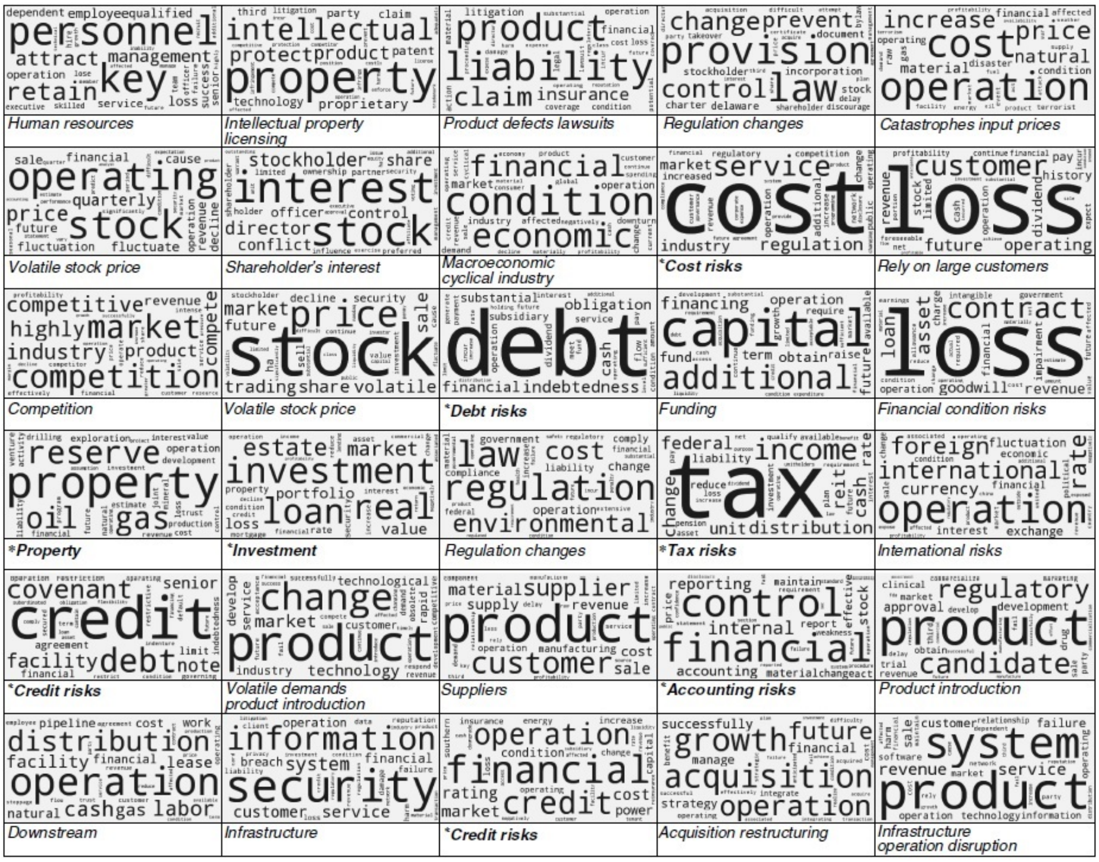


Figure 4.8: Labeling of visualized topics.

Notes: 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 "*". Topic 1 to 30 are displayed from left to right, top to bottom.

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% (498/1327) 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 Section 4.4.1. As shown in Figure 4.8, our learned topics via 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 which 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.

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 which 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 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 comes 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 k-th topic inferred by model m in word intrusion task is defined as the fraction of subjects agreeing with model:

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

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

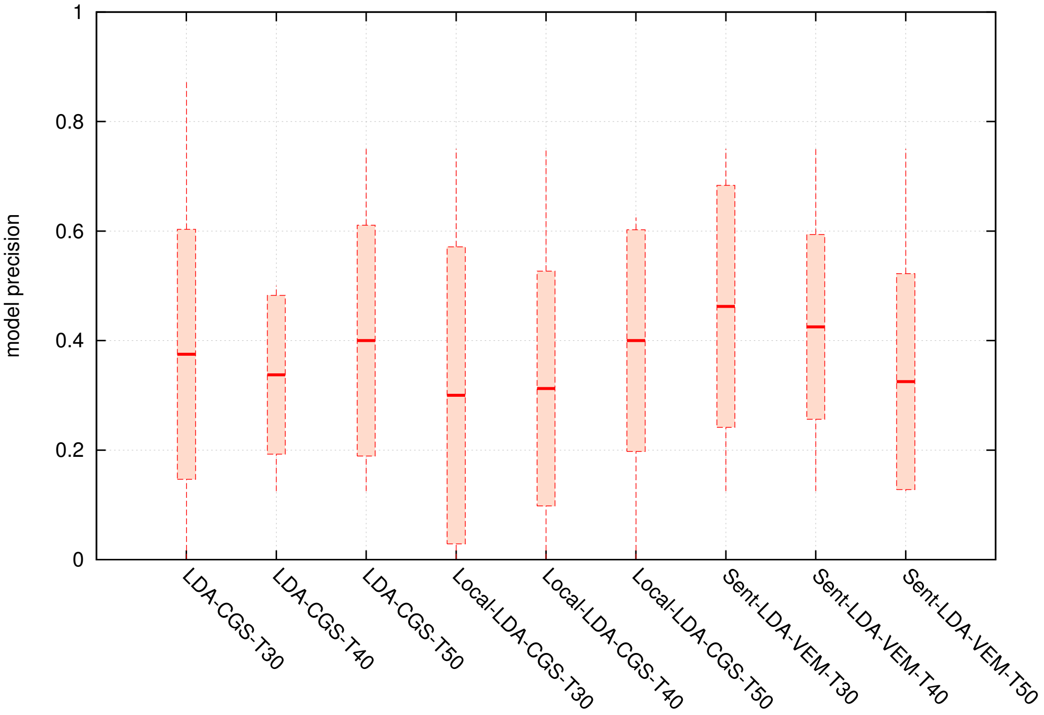


Figure 4.9: Model precision of the word intrusion task.

Figure 4.9 shows boxplots of model precision of three models (LDA-CGS, Local-LDA-CGS and Sent-LDA-VEM) with different number 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 4.3 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 4.3. 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 in Section 4.4.6.

- **Predictive Validation:**

Quinn et al. (2010) and Grimmer (2010) argue that if topics are valid, external events should explain sudden increases in attention to the topics. Following these works, we perform a similar predictive validation using external events.

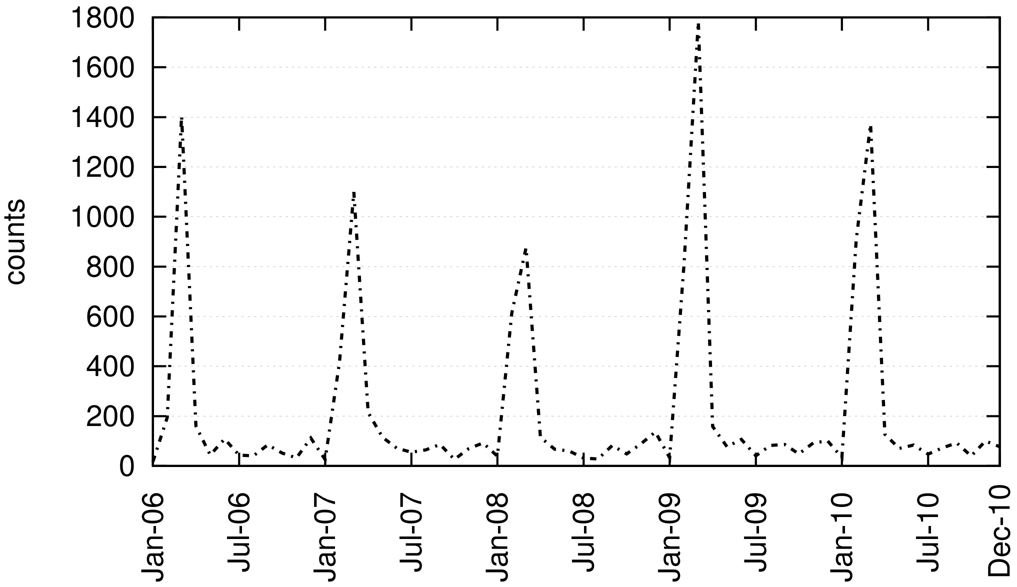


Figure 4.10: Predictive validation of the topic “Macroeconomics risks”.

Figure 4.10 plots the number of risk factors (sentences) released each month about “Macroeconomics risks” inferred by our Sent-LDA-VEM model. The count of risk factors doubles around the year 2009, probably due to 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 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) in documents. Consequently, it is also important to test whether the association between a document and a topic makes sense.

In order 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 4.5. 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 4.8. One observation is that each risk factor indeed discusses only one risk type (topic), which again confirms our key intuition 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 conduct a small-scale qualitative validation of topic assignments for our Sent-LDA-VEM model with 30 topics. The complete topic assignments for risk factors in our dataset are visualized in our publicly available system [4]. Later, we will also present the quantitative validation of topic assignments when comparing our models with the supervised ones.

[4] http://www.comp.nus.edu.sg/~baoyang/10kslda/browse/topic-list.html

Table 4.5: 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.

4.4.5 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 due to 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 is a prerequisite for learning and thus can be utilized when validating. In order to alleviate this skepticism and conduct an equally valid evaluation of our unsupervised method, we construct a training set as in Section 4.4.1 and use it as the ground-truth data for the task of risk factor classification.

In order 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 pre-defined 25 risk types as shown in Figure 4.8. After mapping, we can easily classify the risk factors into risk types 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 (CKNN) algorithm proposed by Huang and Li (2011) for risk factor classification.

Table 4.6 shows the 5-fold cross-validation classification accuracy of the supervised CKNN method with different number of neighbors k, and our Sent-LDA-VEM with number of topics $T = 30$. As can be seen, CKNN performs best when $k = 5$, we thus 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 Grimmer and Stewart (2013)'s observation, that the validation of unsupervised methods in an supervised way *does not* obviate

Table 4.6: 5-folded cross validation classification accuracy.

	CKNN k=2	**CKNN** k=5	**CKNN** k=10	**CKNN** k=15
Mean	0.8130	0.8362	0.8308	0.8233
Std	(± 0.0155)	(± 0.0216)	(± 0.0294)	(± 0.0170)
t-value	1.9507	-	0.3331	1.0517
p-value	0.0869	-	0.7476	0.3237
	CKNN k=20	**LDA-CGS** T=30	**LLDA-CGS** T=30	**SLDA-VEM** T=30
Mean	0.8308	0.0520	0.6060	0.8255
Std	(± 0.0145)	(± 0.0088)	(± 0.0285)	(± 0.0134)
t-value	0.5017	75.2459	14.3963	0.9457
p-value	0.6294	0.0000	0.0000	0.3720

the need for unsupervised methods. This kind of validation is possible only after the unsupervised methods suggest 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.

4.4.6 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 (e.g., perplexity, silhouette coefficient), because 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) noticed this issue and argued 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 the statistical fit (i.e., perplexity as shown in Figure 4.3) and the substantive fit (i.e., semantic coherence as shown in Figure 4.9). 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 4.3 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 4.9. Finally, we choose the number of topics to be 30 at which the model performance in word intrusion tasks is demonstrably the best.

4.4.7 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 in each iteration of the 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 as 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 LDA-VEM model since S will be definitely much smaller than N. In particular, in the same Linux system with dual 3.00GHz CPU and 4.0GB memory, to train a model with 30 topics against our dataset, 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 LDA-VEM algorithm [5]. More importantly, as shown in Figure 4.4 and 4.6, our Sent-LDA-VEM method converges quickly to a much lower perplexity than the LDA-VEM and all the other benchmark models.

[5] We use the implementation of variational EM for LDA available at: http://www.cs.princeton.edu/~blei/lda-c/index.html

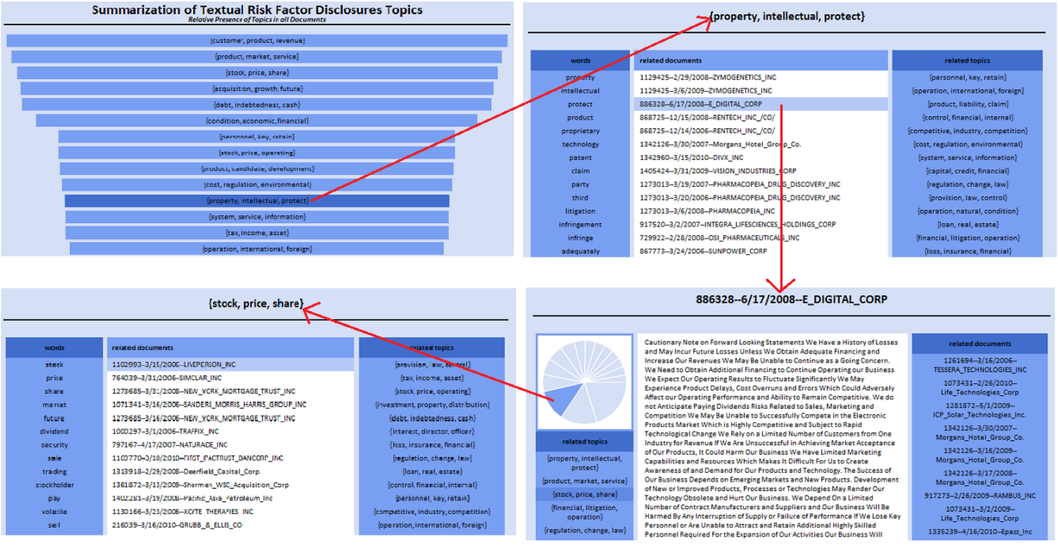
4.5 A Browser of Textual Risk Disclosures

Here, we present the visualization of the outputs of our learned Sent-LDA-VEM model with 30 topics. The visualized model could serve as a browser of corporate risk disclosures [6]. To visualize our model, we adapt the TMVE (Topic Model Visualization Engine) (Chaney and Blei, 2012) which is originally developed for LDA model.

The browser visualizes the topic distributions per document, topic distributions per term, term distributions per topic, list of terms per topic, and relative presence of topics in all documents. For each topic page (top right in Figure 4.11), the most probable terms are listed on the left side in a descending order, and related documents and topics are listed in the middle and right part respectively. For each document page (bottom right in Figure 4.11), the topic proportion is shown on the left side, while the original text and related documents are shown in the middle and right part respectively. For each term page, related terms, documents and topics are listed.

We present an example in Figure 4.11 to demonstrate how to use the browser to navigate the document collection. Beginning in the upper left, we see a set of topics (risk types), each of which is a topic labeled by 3 most probable terms. We click on a topic about "intellectual property protection" and choose a document associated with this topic, which is the risk factor disclosures of a firm called "E Digital" in 2008. The page of this document in the bottom right includes its content and topic proportions. We then explore a related topic about "stock price and shares", which is also discussed in the document. By repeating the process in this example, we can explore more documents in the collection.

[6] The browser is available at: http://www.comp.nus.edu.sg/~baoyang/10kslda/browse/topic-list.html



Summarization of Textual Risk Factor Disclosures Topics

Relative Presence of Topics in all Documents

{customer, product, revenue}
{product, market, service}
{stock, price, share}
{acquisition, growth, future}
{debt, indebtedness, cash}
{condition, economic, financial}
{personnel, key, retain}
{stock, price, operating}
{product, candidate, development}
{cost, regulation, environmental}
{property, intellectual, protect}
{system, service, information}
{tax, income, asset}
{operation, international, foreign}

{property, intellectual, protect}

words	related documents	related topics
property	1129425--2/29/2008--ZYMOGENETICS_INC	{personnel, key, retain}
intellectual	1129425--3/6/2009--ZYMOGENETICS_INC	{operation, international, foreign}
protect	886328--6/17/2008--E_DIGITAL_CORP	{product, liability, claim}
product	868725--12/15/2008--RENTECH_INC_/CO/	{control, financial, internal}
proprietary	868725--12/14/2006--RENTECH_INC_/CO/	{competitive, industry, competition}
technology	1342126--3/30/2007--Morgans_Hotel_Group_Co.	{cost, regulation, environmental}
patent	1342960--3/15/2010--DIVX_INC	{system, service, information}
claim	1405424--3/31/2009--VISION_INDUSTRIES_CORP	{capital, credit, financial}
party	1273013--3/19/2007--PHARMACOPEIA_DRUG_DISCOVERY_INC	{regulation, change, law}
third	1273013--3/20/2006--PHARMACOPEIA_DRUG_DISCOVERY_INC	{provision, law, control}
litigation	1273013--3/6/2008--PHARMACOPEIA_INC	{operation, natural, condition}
infringement	917520--3/2/2007--INTEGRA_LIFESCIENCES_HOLDINGS_CORP	{loan, real, estate}
infringe	729922--2/28/2008--OSI_PHARMACEUTICALS_INC	{financial, litigation, operation}
adequately	867773--3/24/2006--SUNPOWER_CORP	{loss, insurance, financial}

{stock, price, share}

words	related documents	related topics
stock	1102993--3/15/2006--LIVEPERSON_INC	{provision, law, control}
price	764039--3/31/2006--SIMCLAR_INC	{tax, income, asset}
share	1273685--3/31/2008--NEW_YORK_MORTGAGE_TRUST_INC	{stock, price, operating}
market	1071341--3/16/2006--SANDERS_MORRIS_HARRIS_GROUP_INC	{investment, property, distribution}
future	1273685--3/16/2006--NEW_YORK_MORTGAGE_TRUST_INC	{debt, indebtedness, cash}
dividend	1000297--3/1/2006--TRAFFIX_INC	{interest, director, officer}
security	797167--4/17/2007--NATURADE_INC	{loss, insurance, financial}
sale	1100770--3/10/2010--FIRST_PACTRUST_BANCORP_INC	{regulation, change, law}
trading	1313918--2/29/2008--Deerfield_Capital_Corp.	{loan, real, estate}
stockholder	1361872--3/13/2009--Shermen_WSC_Acquisition_Corp	{control, financial, internal}
pay	1402281--3/19/2008--Pacific_Asia_Petroleum_Inc	{personnel, key, retain}
volatile	1130166--3/23/2006--XCYTE_THERAPIES_INC	{competitive, industry, competition}
sell	216039--3/16/2010--GRUBB_&_ELLIS_CO	{operation, international, foreign}

886328--6/17/2008--E_DIGITAL_CORP

Cautionary Note on Forward Looking Statements We Have a History of Losses and May Incur Future Losses Unless We Obtain Adequate Financing and Increase Our Revenues We May Be Unable to Continue as a Going Concern. We Need to Obtain Additional Financing to Continue Operating our Business We Expect Our Operating Results to Fluctuate Significantly We May Experience Product Delays, Cost Overruns and Errors Which Could Adversely Affect our Operating Performance and Ability to Remain Competitive. We do not Anticipate Paying Dividends Risks Related to Sales, Marketing and Competition We May Be Unable to Successfully Compete in the Electronic Products Market Which is Highly Competitive and Subject to Rapid Technological Change We Rely on a Limited Number of Customers from One Industry for Revenue If We Are Unsuccessful in Achieving Market Acceptance of Our Products, It Could Harm Our Business We Have Limited Marketing Capabilities and Resources Which Makes It Difficult For Us to Create Awareness of and Demand for Our Products and Technology. The Success of Our Business Depends on Emerging Markets and New Products. Development of New or Improved Products, Processes or Technologies May Render Our Technology Obsolete and Hurt Our Business. We Depend On a Limited Number of Contract Manufacturers and Suppliers and Our Business Will Be Harmed By Any Interruption of Supply or Failure of Performance If We Lose Key Personnel or Are Unable to Attract and Retain Additional Highly Skilled Personnel Required For the Expansion of Our Activities Our Business Will

related topics
{property, intellectual, protect}
{product, market, service}
{stock, price, share}
{financial, litigation, operation}
{regulation, change, law}

related documents
1261694--3/16/2006--TESSERA_TECHNOLOGIES_INC
1073431--2/26/2010--Life_Technologies_Corp
1281872--5/1/2009--ICP_Solar_Technologies_Inc.
1342126--3/30/2007--Morgans_Hotel_Group_Co.
1342126--3/16/2009--Morgans_Hotel_Group_Co.
1342126--3/17/2008--Morgans_Hotel_Group_Co.
917273--2/26/2009--RAMBUS_INC
1073431--3/2/2009--Life_Technologies_Corp
1335239--4/16/2010--Epazz_Inc

Figure 4.11: A browser of textual risk disclosures.

4.6 Summary

In this chapter, we study the problem of extracting individual risk types from risk disclosures without pre-defined them. To solve this problem, we propose an extended LDA topic model, called Sent-LDA, and its learning algorithm. Based on our "one-sentence-per-topic" observation, our model incorporates the additional information of sentence structure by assuming that each sentence is generated by only one topic. To demonstrate the effectiveness of our proposed model, we conduct experiments to evaluate both the statistical fit (measured by conventional metrics including perplexity and silhouette coefficient) and the substantive fit (i.e., the quality of discovered information measured by human judgment). We show that our proposed model (i.e., Sent-LDA model coupled with variational EM learning algorithm) outperforms all competing unsupervised methods, and could find more meaningful topics for representing risk types. We also show that our proposed unsupervised model performs equally well with supervised method, but could reduce the amount of human effort to a large extent by estimating rather than pre-defining risk types. We further visualize the outputs of our learned model, which could serve as a browser facilitating the navigation of large amount of textual risk disclosures.

Chapter 5

Market Reactions to Individual Risk Types

In this chapter, we continue the analysis of extracted risk types in the previous study in Chapter 4. In particular, we conduct an empirical study to investigate whether and how individual risk types in corporate risk disclosures will affect the post-disclosure risk perceptions of investors.

5.1 Overview

Corporate disclosure is an important way for management to communicate firm performance and governance to various stakeholders, especially outside investors, and is critical to the functioning of an efficient capital market (Healy and Palepu, 2001). Specifically, it is believed that adequate corporate disclosures could enhance the information reflected in stock price in the sense that they reduce the information asymmetry between outside investors and informed market participants like company management (Healy and Palepu, 2001). Therefore, they will result in many desirable consequences, including the efficient allocation of resources in an economy, capital market development, liquidity in the market, decreased cost of capital, lower return volatility, and high analyst forecast accuracy (Diamond and

Verrecchia, 1991; Healy and Palepu, 2001; Bushman and Smith, 2001; Core, 2001; Easley et al., 2002; Easley and O'hara, 2004; Lambert et al., 2007).

Regulators, such as SEC, have also realized the importance of the corporate disclosures, and believed that investors will benefit from the disclosures about the risks and uncertainties of firms. Beginning 2005, the SEC mandated firms to include a "risk factor" section in their 10-K forms to discuss "the most significant factors that make the company speculative or risky". Despite this effort, whether corporate risk disclosures, especially those in this newly-created risk disclosure section, are truly informative to investors remains an open empirical question (Kravet and Muslu, 2013; Campbell et al., 2014). Specifically, there are competing arguments about whether and how risk disclosures will affect investors' 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 investors' 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 (2013) 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) find 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 previous mixed evidence is due to the assumed unidirectional relation between risk disclosures and the measures of market reactions (e.g., cost of capital, stock return volatility). They thus hypothesize that the disclosure tone will affect the direction of the relation, and test a directional relation – favorable disclosures will result in lower return volatility while unfavorable disclosures will lead to higher return volatility.

In this study, we examine the market (investors) reactions to the corporate risk disclosures. We also believe that the relation between disclosures and market reactions is directional. But different from Kothari et al. (2009), we hypothesize that the direction of the relation depends on the semantic content of disclosures, i.e., the individual risk types extracted using the proposed method in the previous study in Chapter 4.

The rest of this chapter is organized as follows. Section 5.2 presents the research question and our hypothesis. Section 5.3 describes the data preparation, including sample selection, variable description and sample statistics. Section 5.4 presents our econometric model specification, its estimation results, and some tests of the explanatory power of risk type variables. Section 5.5 discusses the main findings and their implications. Finally, Section 5.6 provides a brief summary of the study in this chapter.

5.2 Research Question and Hypothesis

In this study, we examine the information content of corporate risk disclosures in the newly-created "risk factor" section in 10-K forms. Specifically, our research question is: *whether and how individual risk types in textual disclosures will affect the post-disclosure risk perceptions of investors.*

Textual risk disclosures present investors with firms' assessments about future contingencies, and they differ from other corporate disclosures in that they guide investors about the range of future performance rather than the level of future performance (Kravet and Muslu, 2013). 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 the firms' future performance.

Besides, existing literature suggests a unidirectional relation between disclosures and market reactions (Kravet and Muslu, 2013). We move the literature forward by testing directional links – investors' risk perception will depend on the specific risk types disclosed. We limit our analysis to the newly-created risk disclosure section, 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 the disclosure tone.

5.3 Data Preparation

In this section, we describe the data preparation for our empirical study on the effects of risk disclosures on investors' risk perceptions.

Our initial sample includes all 10-K forms collected from 2006 to 2010 in the previous study as described in Section 4.4.1. 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 3-year observations.

The main variables in our final sample are summarized below. The summary statistics of these variables are shown in Table 5.1.

- Dependent variable
  - $SRVA_{it}$: firm i's stock return volatility during the first two months after the filing of disclosures in year t

  The dependent variable of interest in our study is the *post-disclosure risk perceptions* of investors. This variable is measured using *stock return volatility*, which is a prominent proxy for diverging investor opinions in the finance literature (Shalen, 1993; Garfinkel, 2009). Following Kravet and Muslu (2013), 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 post-disclosure stock return volatility.

Table 5.1: Summary statistics.

Variable	Mean	Std. Dev.	Min	Max
logSRVA (post-disclosure risk perception)	-7.193569	1.309828	-12.91094	2.397478
logSRVB (pre-disclosure risk perception)	-7.138248	1.236654	-12.0007	1.37464
Price (stock price)	23.099	30.7174	0.06	767.5
logSize (market value of equity)	20.14738	1.89597	13.25745	26.14966
logTrd (trading volume)	12.34596	2.480582	0	19.97997
Eps (earning per share)	0.7405169	6.935572	493.73	110.36
Wc (word count)	137.432	71.29546	0	556
topic1 (human resources risks)	0.8221123	0.7065545	0	5
topic2 (intellectual property & licensing risks)	0.8464644	1.225116	0	9
topic3 (product defects lawsuits)	0.82081	0.9561522	0	13
topic4 (regulation changes: legislation)	0.4952468	0.8091989	0	9
topic5 (catastrophes & input prices)	0.8669098	1.410339	0	19
topic6 (volatile stock price)	0.6322438	0.8943585	0	9
topic7 (shareholder’s interest)	0.6017711	1.371396	0	17
topic8 (macroeconomic & cyclical industry risks)	0.9795546	1.216999	0	15
topic9 (cost risks)	0.430655	1.343795	0	29
topic10 (rely on large customers)	0.5349655	0.8099056	0	6
topic11 (competition risks)	1.008986	0.9760335	0	9
topic12 (volatile stock price)	0.8441203	1.191728	0	12
topic13 (debt risks)	0.5012371	0.9245951	0	11
topic14 (funding risks)	0.6129704	0.8844219	0	9
topic15 (financial condition risks)	0.7318661	1.382401	0	16
topic16 (property risks)	0.4777966	1.566448	0	16
topic17 (investment risks)	0.7330382	2.296263	0	42
topic18 (regulation changes: environment)	1.174632	1.15501	0	14
topic19 (tax risks)	0.7092069	1.996208	0	23
topic20 (international risks)	0.8910014	1.065195	0	11
topic21 (credit risks)	0.415028	0.8383478	0	13
topic22 (volatile demands & production introduction)	0.6799062	0.9849882	0	18
topic23 (supplier risks)	1.039849	1.521993	0	14
topic24 (accounting risks)	0.4309155	0.7761337	0	11
topic25 (production introduction)	1.164344	3.636352	0	33
topic26 (downstream risks)	0.4211486	1.233909	0	30
topic27 (infrastructure risks)	0.5014976	1.160349	0	12
topic28 (credit risks)	0.5037114	1.28771	0	20
topic29 (acquisition & restructuring risks)	1.102748	1.144197	0	15
topic30 (infrastructure & operation disruption)	1.016148	2.006242	0	21

Notes: The number of observations is 7679. Column “Mean”, “Std. Dev.”, “Min”, “Max means” represent the mean, standard deviation, minimum value, and maximum value of corresponding variables in the sample.

- Independent variables of interest
  - *RiskDisclosure*: individual risk types in disclosures

  The independent variables of interest in our study are the individual risk types contained in risk disclosures. These risk type variables are extracted in the previous study in Chapter 4. Specifically, we extract 30 individual risk type variables using our proposed *Sent-LDA-VEM* model (with 30 topics). Risk types are aggregated at the document level, and each document (i.e., "risk factor" disclosure section in a 10-K form) is a firm-year observation. That is, each risk disclosure document will be quantified into a vector with 30 dimensions, and each dimension corresponds to a risk type. The extracted 30 risk types are shown in Figure 4.8 in the previous chapter. For convenience, we index them using topic ids, as shown in Table 5.1.

- Control variables

  To obtain unbiased estimates of the effects of individual risk types on investors' risk perceptions, we control for important relevant factors suggested by prior literature (Kravet and Muslu, 2013; Campbell et al., 2014) as follows.
  - $SRVB_{it}$: firm i's stock return volatility during the two months before the filing of disclosures in year t;
  - $Price_{it}$: firm i's closing stock price at the filing day of risk disclosures in year t
  - $LogSize_{it}$: the log of firm i's market value of equity at the filing day of risk disclosures in year t
  - $LogTrd_{it}$: the log of firm i's trading volume at the filing day of risk disclosures in year t

- Eps_{it}: firm i's earning per share at the filing day of risk disclosures in year t
- Wc_{it}: the word counts of textual risk disclosures of firm i in year t
- *Dummies*: a set of time dummies at the yearly level, and a set of industry dummies based on firms' standard industrial classification code

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 5.1. 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 gray 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.

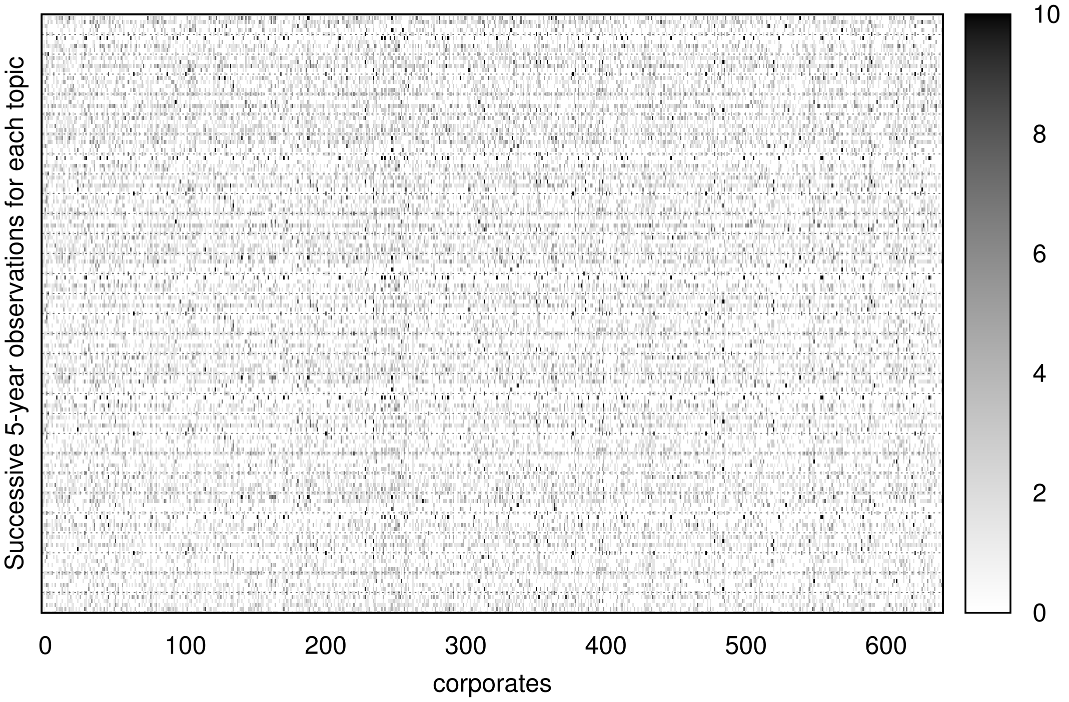
5.4 Econometric Model

In this section, we present the econometric model for our empirical study, including its model specification and estimation results.

5.4.1 Model Specification

We model the influence of individual risk types on the post-disclosure risk perceptions of investors. We estimate a fixed effects linear panel model as shown in

Figure 5.1: Heatmap of the count matrix of risk types.



Equation 5.1.

$$logSRVA_{it} = \alpha_i + \beta_1 \cdot logSRVB_{it} + \beta_2 \cdot Price_{it} + \beta_3 \cdot logSize_{it} + \beta_4 \cdot logTrd_{it} + \beta_5 \cdot Eps_{it} + \beta_6 \cdot Wc_{it} + \beta_T \cdot RiskDisclosure_{it} + \varepsilon_{it} \tag{5.1}$$

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, $logSRVA_{it}$ is the dependent variable, i.e., post-disclosure risk perceptions of investors. The independent variables of interest are 30 individual risk types extracted, denoted as $RiskDisclosure_{it}$. Their corresponding coefficients β_T interpret the influence of risk types on the dependent variable. To obtain unbiased estimates of the effect of risk disclosures on investors' risk perceptions, 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) $LogSize_{it}$, the log of firm i's market value of equity at the filing day of risk disclosures in year t; (4) $LogTrd_{it}$, 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; (7) a set of time dummies at the yearly level and a set of industry dummies based on firms' standard industrial classification code.

5.4.2 Estimation Results

We first estimate a fixed effects (FE) model of investors' post-disclosure risk perceptions ($SRVA$) on all control variables. This baseline model is presented in column (3) of Table 5.2. As can be seen, the control variables have some explanatory power and their coefficients have the expected signs. Specifically, investors' pre-disclosure risk perceptions $SRVB$, the log of firms' size ($LogSize$) and the log of firms' trading volume ($LogTrd$) are significantly associated (positively, negatively and positively respectively) with post-disclosure risk perception. More importantly, we find that the word counts of risk disclosures are significantly and positively associated with the post-disclosure risk perception. Particularly, an additional unique word is associated with a 0.11% increase in investors' post-disclosure risk perception. This finding is consistent with the previous studies in (Campbell et al., 2014; Kravet and Muslu, 2013).

We next estimate a full fixed effects (FE) model by further including all risk types. This full model is presented in column (1) of Table 5.2. We choose the FE model rather than a random effects (RE) model because the Hausman test suggests that the RE estimates are inconsistent ($\chi^2 = 221.89, p = 0.000$).

Table 5.2: Estimation results.

Variable	(1) FE	(2) RE	(3) FE-Controls
topic1	-0.0623** (0.050)	-0.0458*** (0.006)	
topic2	-0.0399 (0.185)	0.0069 (0.588)	
topic3	0.0108 (0.629)	0.0106 (0.378)	
topic4	-0.0826*** (0.006)	-0.0344** (0.02)	
topic5	-0.0208 (0.261)	-0.0145 (0.125)	
topic6	-0.0120 (0.627)	-0.0050 (0.698)	
topic7	0.0181 (0.295)	0.0119 (0.223)	
topic8	0.0295** (0.039)	0.0226** (0.014)	
topic9	0.0312 (0.113)	0.0208** (0.028)	
topic10	-0.0155 (0.545)	-0.0087 (0.536)	
topic11	-0.0255 (0.263)	-0.0265** (0.020)	
topic12	0.0189 (0.342)	0.0096 (0.354)	
topic13	0.0075 (0.748)	0.0080 (0.520)	
topic14	0.0551** (0.015)	0.0326** (0.012)	
topic15	-0.0217 (0.241)	-0.0087 (0.336)	
topic16	-0.0118 (0.626)	0.0213* (0.055)	
topic17	0.0107 (0.401)	0.0147** (0.014)	
topic18	-0.0364* (0.061)	-0.0251** (0.012)	
topic19	-0.0134 (0.367)	-0.0241*** (0.002)	
topic20	0.0126 (0.563)	-0.0024 (0.825)	
topic21	0.0026 (0.911)	-0.0131 (0.345)	
topic22	-0.0106 (0.646)	0.0014 (0.910)	
topic23	0.0024 (0.897)	-0.0014 (0.877)	
topic24	-0.0039 (0.860)	0.0005 (0.971)	
topic25	0.0167 (0.203)	0.0035 (0.507)	
topic26	-0.0250 (0.275)	-0.0119 (0.226)	
topic27	-0.0410* (0.056)	-0.0147 (0.164)	
topic28	0.0578*** (0.002)	0.0203** (0.044)	
topic29	-0.0108 (0.487)	-0.0224** (0.018)	
topic30	-0.0279* (0.066)	-0.0105 (0.135)	
LogSRVB	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)
LogSize	-0.4415*** (0.000)	-0.2260*** (0.000)	-0.4513*** (0.000)
LogTrd	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	$\chi^2 = 211.89, p = 0.000$		

Notes: p-values in parentheses. *** significant at 1% level; ** significant at 5% level; * significant at 10% level.

Table 5.3: BIC differences between OLS models using risk type variables inferred by different models.

	BIC	BIC Difference with (3)	Evidence
(1) LDA-CGS	-7254.387	8.036	strong support for (3)
(2) Local-LDA-CGS	-7257.063	5.360	positive support for (3)
(3) Sent-LDA-VEM	-7262.423	-	-

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).

5.4.3 Explanatory Power of Risk Type Variables

To test the joint significance of our discovered risk type variables (topic1 to topic30), we conduct a likelihood ratio test on the nested models "FE full" and "FE Controls" in column (1) and (3) of Table 5.2. 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 demonstrated in Chapter 4 that our Sent-LDA-VEM model could find incremental information, i.e., risk types that are ignored by (Huang and Li, 2011) when pre-defining the categories for supervised learning methods. In order to test the joint significance of our additional risk types, we conduct another likelihood ratio test on two nested models – "FE full" model and "FE full" model excluding the 8 risk types (i.e., the new risk types preceded by asterisk as shown in Figure 4.8, which are not found 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.

As shown in the previous study in Chapter 4, original LDA model and Local-LDA model 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 BIC (Bayesian Information Criterion) statistic 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 non-nested models (Raftery, 1995). In particular, we run the OLS (ordinary least squares) model using the same dependent and independent variables listed in column (1) of Table 5.2. 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 5.3 reports the BIC model fit for all three topic models. According to the guidelines stipulated by 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.

5.5 Findings and Implications

In this section, we discuss our main findings about the effects of individual risk types on the post-disclosure risk perceptions of investors. Our discussion is based on the estimation results of the full FE model as shown in column (1) in Table 5.2. Interestingly, our findings provide support for all three competing arguments about whether and how risk disclosures will affect investors' risk perceptions (as introduced in Section 5.1).

Support for Null Argument

First, we find that 22 out of 30 risk types have no significant influence on the post-disclosure 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 in risk disclosures, and warned firms to "avoid generic risk factor disclosures that could apply to any company". To examine whether the effort of SEC has 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, 2013). One limitation of these studies is that they cannot drill down into analyzing fine-grained risk types due to 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, topic11 (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 topic6 and topic12 (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. While 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 topic8 (Macroeconomic risks), topic14 (Funding risks) and topic28 (Credit risks), are positively associated with the post-disclosure 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 post-disclosure risk perceptions at 5%, 5%, 1% significant level 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 post-disclosure 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 which may be compounded by liquidity risk) are informative and will increase the post-disclosure 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, 2013) which 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 post-disclosure 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 (2013) 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 topic1 (Human resources risks), topic4 (Regulation changes: shareholders' interests), topic18 (Regulation changes: environment), topic27 (Infrastructure risks: information security), and topic30 (Infrastructure risks: disruption), are negatively associated with the post-disclosure 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 post-disclosure 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 post-disclosure 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 post-disclosure 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 which cannot be diversified. Our results suggest that the informative disclosure of certain unsystematic risks will even decrease the post-disclosure 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, 2013) which 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.

Implications

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' post-disclosure 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 while the rest two thirds lack informativeness. This challenges the findings of prior studies (Campbell et al., 2014) that the disclosures in newly added section 1A of

10-K forms are 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.

5.6 Summary

In this chapter, we continue the analysis of extracted risk types in the previous study in Chapter 4. In particular, we conduct an empirical study to investigate the effects of individual risk types on the post-disclosure risk perceptions of investors. Different from prior works, our empirical study provides 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. Specifically, we find that: (1) around two thirds of risk types have no significant influence on the post-disclosure risk perceptions of investors, lending support for the null argument that risk disclosures are by and large boilerplate; (2) the disclosure of 3 types of financial and systematic risks will increase the post-disclosure risk perceptions of investors, lending support for the divergence argument; and (3) the disclosure of 5 types of legal and idiosyncratic risks will decrease the post-disclosure risk perceptions of investors, lending support for the convergence argument. Our findings have implications for both managers and regulators.

Chapter 6

Conclusion

In this chapter, we provide concluding remarks for the thesis. Specifically, we summarize each study and recap the corresponding contributions. We then identify the limitations of the studies, and finally discuss the possible directions for future research.

6.1 Concluding Remarks

This thesis is comprised of three studies. First, we study the problem of cross-domain text classification (in general), and cross-industry risk sentiment analysis (in particular). To solve this problem, we propose an extended LDA topic model, called PSCCLDA, and its learning algorithm. With the purpose of overcoming two observed limitations of existing methods, our proposed model could explicitly distinguish the domain-independent and domain-specific topics by resorting to the cross-collection topic models, and exploit the label information for inferring more predictive topics by embedding the supervised logistic regression model. Experimental results on nine standard dataests demonstrate the effectiveness of our model for cross-domain text classification in general, and its superiority over the state-of-the-art cross-domain learning methods. Experiential results of 42 tasks for cross-industry risk sentiment on MD&A disclosures in 10-K forms show the effectiveness of our proposed model, and its superiority over existing supervised learning methods that have been adopted in the financial accounting domain.

Second, we study the problem of extracting individual risk types from risk disclosures without pre-defined them. To solve this problem, we propose an extended LDA topic model, called Sent-LDA, and its learning algorithm. Based on our "one-sentence-per-topic" observation, our model incorporates the additional information of sentence structure by assuming that each sentence is generated by only one topic. To demonstrate the effectiveness of our proposed model, we conduct experiments to evaluate both the statistical fit (measured by conventional metrics including perplexity and silhouette coefficient) and the substantive fit (i.e., the quality of discovered information measured by human judgment). We show that our proposed model (i.e., Sent-LDA model coupled with variational EM learning algorithm) outperforms all competing unsupervised methods, and could find more meaningful topics for representing risk types. We also show that our proposed unsupervised model performs equally well as supervised methods, but could reduce the amount of human effort to a large extent by estimating rather than pre-defining risk types. We further visualize the outputs of our learned model, which could serve as a browser facilitating the navigation of large amount of textual risk disclosures.

Third, we continue the analysis of extracted risk types in the previous study. In particular, we conduct an empirical study to investigate the effects of individual risk types on the post-disclosure risk perceptions of investors. Different from prior works, our empirical study provides 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. Specifically, we find that: (1) around two thirds of risk types have no significant influence on the post-disclosure risk perceptions of investors, lending support for the null argument that risk disclosures are by and large boilerplate; (2) the disclosure of 3 types of financial and systematic risks will increase the post-disclosure risk perceptions of investors, lending support for the divergence argument; and (3) the disclosure of 5 types of legal and idiosyncratic risks will decrease the post-disclosure risk perceptions of investors, lending support for the convergence argument. Our findings have implications for both managers and regulators.

In summary, the main contribution of this thesis is the development of two variants of LDA topic model for identifying risk sentiment and extracting various risk types from textual risk disclosures. The proposed methods can facilitate the analysis of corporate risk disclosures by reducing the amount of manual effort substantially, and are among the first to introduce the cross-domain learning and unsupervised learning methods into the field of financial accounting. Moreover, by taking advantage of the proposed method, an empirical study is enabled to examine the market reactions to risk disclosures at the individual risk type level. The findings reconcile the conflicting arguments on the effects of risk disclosures on post-disclosure risk perceptions of investors in accounting literature.

6.2 Limitations

The studies in this thesis are not without limitations, which are summarized below.

Limitations of Our Model for Cross-Industry Analysis

Our PSCCLDA model for cross-industry analysis is subjected to several limitations. First, our model is limited to binary classification tasks. To model the observed class labels (e.g., risk sentiment labels) of documents in the training source domain, we embed the logistic regression model into the unsupervised cross-collection LDA model. Logistic regression model is an effective generative classifier but can only be applied for binary classification problem. Since we might have multiple risk sentiment labels (e.g., positive, negative and neutral) in some cases, it is more desirable to design a model that can be directly used for multi-class classification problem in the context of cross-domain learning. To this end, one possible solution is to generalize the logistic regression model to the softmax regression model (Wang et al., 2009a), which could be directly applied for the multi-class classification problem (e.g., cross-industry sentiment analysis in our case).

Second, our PSCCLDA model is not non-parametric. There are two important parameters to be set, including the number of topics and the importance weight of collection-independent and collection-dependent topics. Although we have demonstrated that our proposed model is not sensitive to these parameters in Section 3.4, these parameters might need to be carefully tuned when the model is applied for a new problem. In this sense, a non-parametric model is more preferable. One possible solution might be the hierarchical Dirichlet process (Teh et al., 2005) which could automatically infer the number of topics from data.

Third, our PSCCLDA model is actually a general cross-domain text classification algorithm, which is not optimized for the specific task of cross-industry risk sentiment analysis. For better performance, the model should be augmented with existing domain knowledge in the context of risk sentiment analysis. One possible augment might be the inclusion of the prior knowledge of risk sentiment words (Loughran and McDonald, 2011) via the Dirichlet prior as in (Jo and Oh, 2011).

Limitations of Our Model for Extracting Individual Risk Types

Our Sent-LDA model for extracting individual risk types is subjected to several limitations. First, we do not explore the correlations between the learned topics (i.e., risk types). Clearly, some risk types are more related with each other and should be merged together. It would be useful to explore these correlations and perform a hierarchical clustering (Han et al., 2006) of the learned topics so that we can obtain a more concise taxonomy of risk types.

Second, our Sent-LDA model is not non-parametric. There is one important parameter, i.e., the number of topics, that needs to be carefully chosen. Although we have discussed how to set this parameter in Section 4.4.6, it is not an easy task to choose its optimal value. One possible solution might be the hierarchical Dirichlet process (Teh et al., 2005) which could automatically infer the number of topics from data.

Third, our Sent-LDA model does not support the interactive model learning. We have elicited the human judgments for evaluating the quality of discovered information of our model. These human judgments from domain experts are valuable domain knowledge which could be in turn incorporated into the model for enhancement. However, our current model does not provide the mechanism, such as in (Hu et al., 2011), to incorporate expert feedback for adjusting the model leaning in an interactive manner.

Limitations of Our Empirical Study on Effects of Individual Risk Types

Our empirical study on the effects of individual risk types is not without its limitations. First, we only look at risk disclosures from regulated corporate annual reports (i.e., 10-K forms), but do not consider risk disclosures from other sources, such as voluntary communications and information intermediaries. The disclosures from various sources have different levels of *credibility* and *timeliness* (Kothari et al., 2009), which should be taken into account when examining the effects of risk disclosures.

Second, we do not consider the tone (sentiment) of the extracted risk types. Some prior studies (Kothari et al., 2009) provide the evidence that the tone of risk disclosures will affect the risk perceptions of investors. Therefore, the tone, in addition to the risk types, should be taken into account when examining the effects of risk disclosures. To this end, we need to design novel text analysis methods which can simultaneously extract the individual risk types and identify the tone associated with them.

6.3 Directions for Future Work

Apart from the future work for addressing the aforementioned limitations, there are some interesting future research directions as discussed below.

More Robust Evaluation Methods

The evaluation of unsupervised topic models is an interesting direction for future research. The challenge of evaluating topic models lies on their unsupervised nature (Grimmer and Stewart, 2013). When the model is extended to a supervised version, such as our PSCCLDA model proposed in Chapter 3, it is easy to perform evaluation in the same way as the supervised models. But if the model is unsupervised, there usually lacks the ground-truth data for evaluation. Conventionally, unsupervised topic models are evaluated in terms of the statistical model fit, using metrics such as *perplexity* and *empirical likelihood* (Wallach et al., 2009b). Recently, some researchers begin to notice the limitations of the exiting evaluation methods for unsupervised topic models. Particularly, some researchers (Chang et al., 2009; Grimmer and Stewart, 2013) argue that the quality of the discovered semantic structure is more important than the statistical fit when evaluating topic models, especially when the models are employed for social science research. Von Luxburg et al. (2012) also suggest that unsupervised 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/she has never read.

More Scalable Inference Algorithms

Another possible direction for future research is to improve the scalability of the model inference for the LDA and its variants. Existing learning algorithms for the LDA-style models require heavy computations. Due to the prevalence of large datasets, there is a need to design more scalable and efficient learning algorithms. In this direction, there have been some attempts to design the scalable parallel framework for learning the LDA model, such as the Hadoop MapReduce framework in (Ahmed et al., 2012; Zhai et al., 2012). In future, it will be extremely useful to design such scalable framework for the extended topic models, such as those proposed in this thesis.

More Empirical Studies on Individual Risk Types

There are many opportunities to examine the effects of risk disclosures at the individual risk type level. Financial accounting researchers have long recognized the importance of corporate disclosures, and have conducted various empirical studies to examine their effects (Healy and Palepu, 2001). A prerequisite for this line of research is the effective text analysis method for quantifying the variables of interest from text, including amount, tone, and readability (Li, 2010b). However, few studies examine the effects of another important variable, i.e., individual risk type, due to the lack of effective methods for extracting them. Our proposed solution for extracting this variable enables the empirical inactivation of corporate risk disclosures at the individual risk type level. In the thesis, we only examine the effects of risk disclosures on the post-disclosure risk perceptions of investors. In future, there are many other dependent variables that can be investigated, such as future earnings, abnormal returns, accounting frauds, and so on.

Bibliography

Ahmed, A., Aly, M., Gonzalez, J., Narayanamurthy, S., and Smola, A. J. (2012). Scalable inference in latent variable models. In *Proceedings of the 5th International Conference on Web Search and Data Mining*, pages 123–132. ACM.

Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294.

Aral, S., Ipeirotis, P., and Taylor, S. (2011). Content and context: Identifying the impact of qualitative information on consumer choice. In *Proceedings of the 32nd International Conference on Information Systems*, pages 1–9. AIS.

Asuncion, A., Welling, M., Smyth, P., and Teh, Y. (2009). On smoothing and inference for topic models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 27–34. UAI Press.

Azzopardi, L., Girolami, M., and Van Risjbergen, K. (2003). Investigating the relationship between language model perplexity and ir precision-recall measures. In *Proceedings of the 26th International Conference on Research and Development in Informaion Retrieval*, pages 369–370. ACM.

Ben-david, S., Blitzer, J., Crammer, K., and Pereira, F. (2006). Analysis of representations for domain adaptation. In *Neural Information Processing Systems*, pages 137–144.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine* *learning*. Springer New York.

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. and Jordan, M. (2006). Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143.

Blei, D. and Lafferty, J. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120. ACM.

Blei, D. and Lafferty, J. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, 1:17–35.

Blei, D. and McAuliffe, J. (2008). Supervised topic models. In *Neural Information Processing Systems*, pages 121–128.

Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128. ACL.

Brody, S. and Elhadad, N. (2010). An unsupervised aspect-sentiment model for online reviews. In *Proceedings of the 11th Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812. ACL.

Bushman, R. M. and Smith, A. J. (2001). Financial accounting information and corporate governance. *Journal of Accounting and Economics*, 32(1):237–333.

Campbell, J., Chen, H., Dhaliwal, D., Lu, H., and Steele, L. (2014). The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies*, 19(1):396–455.

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

Chaney, A. and Blei, D. (2012). Visualizing topic models. In *Proceedings of the 2012 International AAAI Conference on Social Media and Weblogs*, pages 1–4.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*, pages 288–296.

Chang, Y. and Chien, J. (2009). Latent dirichlet learning for document summarization. In *Proceedings of the 34th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1689–1692. IEEE.

Chuang, J., Manning, C. D., and Heer, J. (2012). Termite: visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 74–77. ACM.

Core, J. E. (2001). A review of the empirical disclosure literature: discussion. *Journal of Accounting and Economics*, 31(1):441–456.

Dai, W., Yang, Q., Xue, G.-R., and Yu, Y. (2007). Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine learning*, pages 193–200. ACM.

Dhillon, I. S. and Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1-2):143–175.

Diamond, D. W. and Verrecchia, R. E. (1991). Disclosure, liquidity, and the cost of capital. *The Journal of Finance*, 46(4):1325–1359.

Doyle, G. and Elkan, C. (2009). Accounting for burstiness in topic models. In

Proceedings of the 26th International Conference on Machine Learning, pages 281–288. ACM.

Du, L., Buntine, W., and Jin, H. (2010). A segmented topic model based on the two-parameter poisson-dirichlet process. *Machine Learning*, 81(1):5–19.

Easley, D., Hvidkjaer, S., and O'hara, M. (2002). Is information risk a determinant of asset returns? *The Journal of Finance*, 57(5):2185–2221.

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

Fama, E. and French, K. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fan, R., Chang, K., Hsieh, C., Wang, X., and Lin, C. (2008). Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

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

Frankel, R. M., Johnson, M. F., and Nelson, K. K. (2002). The relation between auditors' fees for nonaudit services and earnings management. *The Accounting Review*, 77(s-1):71–105.

Garfinkel, J. (2009). Measuring investors' opinion divergence. *Journal of Accounting Research*, 47(5):1317–1348.

Girolami, M. and Kabán, A. (2003). On an equivalence between plsi and lda. In *Proceedings of the 26th International Conference on Research and Development in Informaion Retrieval*, pages 433–434. ACM.

Griffiths, T. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235.

Griffiths, T., Steyvers, M., Blei, D. M., and Tenenbaum, J. B. (2005). Integrating topics and syntax. *Neural Information Processing Systems*, 17:537–544.

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

Grimmer, J. and King, G. (2011). General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, 108(7):2643–2650.

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

Han, J., Kamber, M., and Pei, J. (2006). *Data mining: concepts and techniques.* Morgan kaufmann.

Healy, P. M. and Palepu, K. G. (2001). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. *Journal of Accounting and Economics*, 31(1):405–440.

Heinrich, G. (2005). Parameter estimation for text analysis. Technical report.

Hofmann, T. (1999a). Probabilistic latent semantic analysis. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pages 289–296. Morgan Kaufmann.

Hofmann, T. (1999b). Probabilistic latent semantic indexing. In *Proceedings of* *the 22nd International Conference on Research and Development in Information Retrieval*, pages 50–57. ACM.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196.

Hong, L., Dom, B., Gurumurthy, S., and Tsioutsiouliklis, K. (2011). A time-dependent topic model for multiple text streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 832–840. ACM.

Hopkins, D. J. and King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247.

Hu, Y., Boyd-Graber, J., and Satinoff, B. (2011). Interactive topic modeling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 248–257. ACL.

Huang, K. and Li, Z. (2011). A multilabel text classification algorithm for labeling risk factors in sec form 10-k. *ACM Transactions on Management Information Systems*, 2(3):1–19.

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

Jiang, J. (2008). A literature survey on domain adaptation of statistical classifiers. *Available at: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey.*

Jo, Y. and Oh, A. (2011). Aspect and sentiment unification model for online review analysis. In *Proceedings of the 4th International Conference on Web Search and Data Mining*, pages 815–824. ACM.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. (2009). Predicting risk from financial reports with regression. In *Proceedings of the 2009 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280. ACL.

Kothari, S., Li, X., and Short, J. (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 Review*, 84(5):1639–1670.

Kravet, T. and Muslu, V. (2013). Textual risk disclosures and investors' risk perceptions. *Review of Accounting Studies*, 18(4):1088–1122.

Kupiec, J. (1992). Robust part-of-speech tagging using a hidden markov model. *Computer Speech & Language*, 6(3):225–242.

Lacoste-Julien, S., Sha, F., and Jordan, M. I. (2008). Disclda: discriminative learning for dimensionality reduction and classification. In *Neural Information Processing Systems*, pages 897–904.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.

Lambert, R., Leuz, C., and Verrecchia, R. E. (2007). Accounting information, disclosure, and the cost of capital. *Journal of Accounting Research*, 45(2):385–420.

Li, F. (2010a). The information content of forward-looking statements in corporate filings: a naive bayesian machine learning approach. *Journal of Accounting Research*, 48(5):1049–1102.

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

Li, F. and Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition*, pages 524–531. IEEE.

Li, L., Jin, X., and Long, M. (2012). Topic correlation analysis for cross-domain text classification. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*. AAAI.

Lin, C., He, Y., and Everson, R. (2011). Sentence subjectivity detection with weakly-supervised learning. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 1153–1161.

Lin, C., Weng, R., and Keerthi, S. (2008). Trust region newton method for logistic regression. *The Journal of Machine Learning Research*, 9:627–650.

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

Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining Text Data*, pages 415–463. Springer.

Long, M., Wang, J., Ding, G., Cheng, W., Zhang, X., and Wang, W. (2012). Dual transfer learning. In *Proceedings of the 12th SIAM International Conference on Data Mining*, pages 540–551. SIAM.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Loughran, T. and McDonald, B. (2013). Ipo first-day returns, offer price revisions, volatility, and form s-1 language. *Journal of Financial Economics*, 109(2):307–326.

Lu, B., Ott, M., Cardie, C., and Tsou, B. (2011). Multi-aspect sentiment analysis with topic models. In *Proceedings of the 11th IEEE International Conference on Data Mining Workshops*, pages 81–88. IEEE.

McCallum, A., Nigam, K., et al. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, volume 752, pages 41–48.

Mei, Q., Shen, X., and Zhai, C. (2007). Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 490–499. ACM.

Mimno, D. and McCallum, A. (2008). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 411–418. UAI Press.

Mirakur, Y. (2011). Risk disclosure in sec corporate filings. *Working paper available at: http://repository.upenn.edu/wharton_research_scholars/85*.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2-3):103–134.

O'Connor, B., Bamman, D., and Smith, N. (2011). Computational text analysis for social science: model assumptions and complexity. In *Proceedings of the 2nd Workshop on Computational Social Science*, pages 1–8.

Pan, S., Ni, X., Sun, J., Yang, Q., and Chen, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*, pages 751–760. ACM.

Pan, S. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Paul, M. and Girju, R. (2009). Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1408–1417. ACL.

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st International ACM Conference on Research and Development in Information Retrieval*, pages 275–281. ACM.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., and Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological 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. *The Accounting Review*, 74(3):251–280.

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 248–256. ACL.

Ramage, D., Manning, C. D., and Dumais, S. (2011). Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining*, pages 457–465. ACM.

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

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Schrand, C. M. and Elliott, J. A. (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).

SEC (2014). How to read a 10-k. https://www.sec.gov/answers/reada10k.htm.

Sha, F. and Pereira, F. (2003). Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language*, pages 134–141. ACL.

Shalen, C. (1993). Volume, volatility, and the dispersion of beliefs. *Review of Financial Studies*, 6(2):405–434.

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

Sun, Y., Deng, H., and Han, J. (2012). Probabilistic models for text mining. In *Mining Text Data*, pages 259–295. Springer.

Tausczik, Y. and Pennebaker, J. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2005). Sharing clusters among related groups: Hierarchical dirichlet processes. In *Neural Information Processing Systems*, pages 1385–1392.

Teh, Y., Kurihara, K., and Welling, M. (2008). Collapsed variational inference for hdp. *Neural Information Processing Systems*, 20(20):1481–1488.

Teh, Y., Newman, D., and Welling, M. (2007). A collapsed variational bayesian inference algorithm for latent dirichlet allocation. *Neural Information Processing Systems*, 19:1353–1360.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168.

Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467.

Titov, I. (2011). Domain adaptation by constraining inter-domain variability of latent feature representation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 62–71.

Titov, I. and McDonald, R. (2008). Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web*, pages 111–120. ACM.

Von Luxburg, U., Williamson, R., and Guyon, I. (2012). Clustering: Science or art? *Journal of Machine Learning Research*, 27:65–80.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305.

Wallach, H. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine learning*, pages 977–984. ACM.

Wallach, H., Jensen, S., Dicker, L., and Heller, K. (2010). An alternative prior process for nonparametric clustering. *Journal of Machine Learning Research*, 9:892–899.

Wallach, H., Mimno, D., and McCallum, A. (2009a). Rethinking lda: Why priors matter. *Neural Information Processing Systems*, 22:1973–1981.

Wallach, H., Murray, I., Salakhutdinov, R., and Mimno, D. (2009b). Evaluation methods for topic models. In *Proceedings of the 26th International Conference on Machine Learning*, pages 1105–1112. ACM.

Wang, C., Blei, D., and Li, F. (2009a). Simultaneous image classification and annotation. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1903–1910. IEEE.

Wang, D., Zhu, S., Li, T., and Gong, Y. (2009b). Multi-document summarization using sentence-based topic models. In *Proceedings of the 4th International Joint Conference on Natural Language Processing*, pages 297–300.

Xue, G., Dai, W., Yang, Q., and Yu, Y. (2008). Topic-bridged plsa for cross-domain text classification. In *Proceedings of the 31st International Conference on Research and Development in Information Retrieval*, pages 627–634. ACM.

Yoo, J. and Choi, S. (2009). Probabilistic matrix tri-factorization. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1553–1556. IEEE.

Zhai, C. (2008). Statistical language models for information retrieval. *Synthesis Lectures on Human Language Technologies*, 1(1):1–141.

Zhai, C., Velivelli, A., and Yu, B. (2004). A cross-collection mixture model for comparative text mining. In *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining*, pages 743–748. ACM.

Zhai, K., Boyd-Graber, J., Asadi, N., and Alkhouja, M. (2012). Mr. lda: A flexible large scale topic modeling package using variational inference in mapreduce. In *Proceedings of the 21st International Conference on World Wide Web*, pages 879–888. ACM.

Zhu, J., Ahmed, A., and Xing, E. P. (2009). Medlda: maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th International Conference on Machine Learning*, pages 1257–1264. ACM.

Zhuang, F., Luo, P., Shen, Z., He, Q., Xiong, Y., Shi, Z., and Xiong, H. (2010). Collaborative dual-plsa: mining distinction and commonality across multiple domains for text classification. In *Proceedings of the 19th International Conference on Information and knowledge Management*, pages 359–368. ACM.