Firm Risk Identification Through Topic Analysis of Textual Financial Disclosures

Xiaodi Zhu, Steve Y. Yang, Somayeh Moazeni
Financial Engineering Program
Stevens Institute of Technology
Hoboken, New Jersey, USA
xzhu@stevens.edu, steve.yang@stevens.edu, smoazeni@stevens.edu

***Abstract*—Corporate risk disclosures as part of U.S. public companies' financial reports are mandated by the Securities and Exchange Commission (SEC) since 2005. It provides forward-looking information about companies' future business and potential risks. This study analyzes risk types revealed in these risk disclosures and examines their potential implications on stock returns. Using 16,110 risk disclosures submitted to the SEC from 2011 to 2015, we apply Sentence Latent Dirichlet Allocation (Sent-LDA) model to infer risk types and propose a novel algorithm to match new factors with existing risk types which generates 90% correct matches. We then quantify the impact of different risk factors on the distribution of stock returns using different time windows. We find that common risk factors, such as accounting risk and acquisition risk, have significant effects on both long-term and short-term stock returns. Some other factors only have short-term or long-term effects on stock returns. These findings provide evidence that the companies' self-disclosed risk factors have significant impacts on subsequent stock return volatility and such impacts can be used to predict potential stock change after the public release of the financial risk disclosures.**

I. Introduction

The annual financial reports issued by public companies are important resources for investors, regulators and policy makers to gain insights and detailed information about corporate financial conditions and potential risks. Financial reports submitted to the Securities and Exchange Commission (SEC) are mandatory to most U.S. publicly listed companies, including both structured data and unstructured data. Structured data like financial statements have been widely studied and used in portfolio management to evaluate the potential risk and stock return of a single company [1], [2]. Recently, unstructured data have become one of the most analyzed source data in academic research and have received more attention from the market [3], [4]. In 2005, the SEC required all filers to discuss company's risk factors as one additional section (Section 1A) in their annual report (10-K) [5]. In 2010, to improve the informativeness of risk disclosures, the SEC issued comment letters asking filers only include specific risk factors related to the individual company [6]. Unlike the structured data that primarily summarizes the past performance of a company, the risk disclosures includes forward-looking information that may reveal the company's future risk and can be potentially used to predict equity risk in investment decisions and portfolio management.

Contextual information in financial narratives is rich enough to convey complex and multifaceted information about company's financial conditions intertwined with future prospects and emerging risks [7]. Previous studies have raised many interests in analyzing risk disclosure data, such as disclosure length, tone, and readability [8]–[10]. Analyzing actual content of risk disclosures, such as risk types, has drawn less attention so far. Discovering different impacts from risk factors can help investors evaluate stock risk better and get more accurate prediction on portfolio performances. A recent study analyzed risk disclosures from 2006 to 2010 and found that 8 risk factors have significant impact on post-disclosure stock volatilities [3].

This study focuses on the analysis of risk types from these risk disclosures and their potential predictive power on the stock returns. First, we adopt Sentence Latent Dirichlet Allocation (Sent-LDA) model [3] to identify risk factors from risk disclosures. We then develop a multi-factor model for three return characteristic variables, i.e. stock return volatility, kurtosis and skewness, which are associated to the risk of companies' stock returns. We also investigate the time effects of risk factors on the post-disclosure stock returns. Moreover, we propose a novel topic matching algorithm combining similarity measure and clustering method to match factors from new financial disclosures to existing risk types which potentially reduces the subjective bias while assigning risk factors to the correct topics. We use risk disclosure in annual financial reports from 2011 to 2015, including 16,110 financial filings from 4,919 unique companies. Our analysis shows that common risk factors, such as acquisition risk and accounting risk, have significant effects on both long-term and short-term stock returns. Some other factors such as down-stream risk, only have short-term effect, while factors including cost/supply risk have long-term effect on stock returns. Our findings provide evidence that the self-disclosed risk factors have significant impact on stock returns and such impact can be used to predict potential stock change after the public release of the financial risk disclosures.

The main contributions of the paper are:

- We propose novel topic matching algorithm to potentially reduce the subjective bias from human judgment while matching risk factors from new disclosure to existing risk types.
- We use risk disclosures from 2011 to 2015 after the SEC

released comment letter to improve the informativeness of risk disclosures [6]. The new data might provide more accurate results on the effects of stock returns.
- We not only study the impact of risk factors on stock return volatility, kurtosis and skewness, but also analyze their time effects by applying multi-factor models on different time windows.

The rest of the paper is organized as follows. Related literatures are reviewed in section II. Section III describes the data and our information extraction approach. In section IV, we discuss empirical results of both topic matching and factors modeling of stock returns. The final section concludes our findings and lays out a further work plan.

II. Literature Review

Unlike structured data, information in financial narratives may provide a potential for information users to predict the company's future performance associated with these innate risks. Earlier studies have provided evidence on the relationship between financial narrative data and corporate risk that financial narratives can provide more information about company's potential risk to investors [11], [12]. To use financial narratives in a quantitative analysis, one has to be able to quantify textual data from financial disclosures. Readability analysis and tone analysis [8], [13] have been widely applied to financial textual data, such as Management Discussion and Analysis (MD&A) [14] and public news [15]. As a result, the long and complex financial reports have been found negatively associated with the performance of company where managers tend to use complicated statement to hide company's actual performance, especially with negative news [16], [17].

The informativeness of the risk factor section was criticized from the beginning, because companies do not need to estimate and quantify the effect of risks and the managers may include all potential risks in the disclosure [4]. The lack of proper risk information may reduce the usefulness of section 1A to information users such as investors [18]. After the issuance of the comment letter by the SEC in 2010 [6], the risk disclosure should start to provide more risk information to investors which can be used for better securities investment decision. [9] found that the increase of risk disclosure have positive association with stock market volatility and the use of risk disclosure increases investors' risk perception. [4] used proportion of certain key words in risk disclosures and showed that the risk disclosure is informative and is able to decreases information asymmetry.

However, quantifying the actual content in risk disclosure, such as risk types, has been less studied. [19] first considered risk types disclosed in financial reports while classifying words in MD&A section. The study shows that company's conditions (i.e. acquisition activity, debt condition) potentially affect the number and length of risk factors disclosed in financial report. [20] applied a supervised classification algorithm on words in Section 1A which assigned multiple risk factors on each disclosure. The study proposed 25 risk types extracted from risk disclosures, which laid the foundation for further risk factor analysis. Although [20] considers different types of risk, the supervised algorithm requires human efforts to identify and predefine a comprehensive list of risk factors in financial disclosures. There are only a few studies analyzing financial narrative contents using unsupervised machine learning method. [21] first applied unsupervised model, Latent Dirichlet Allocation(LDA), on stock recommendations to measure stock selection bias. [22] improved traditional LDA model to Sent-LDA model which assign only one single topic on words in the same sentence. A further study from the same authors demonstrated that the Sent-LDA model is more optimized than traditional LDA model [3]. It also extended 25 risk types proposed in [20] to 30 risk factors and found 8 factors associated with post-disclosure stock market volatility [3].

In contrast to the previous studies, we use latest data in risk disclosures from 2011 to 2015 to avoid the potential effects from financial crisis in 2008. More common risk factors might be generated from the new dataset which are different from factors found in [3]. Moreover, we improve the application of topic model in portfolio management process by proposing a novel algorithm to match risk factors from different models based on word probability distribution. The algorithm eliminates manual selection and reduces subjective bias in the topic matching process. In additional to the study of risk factor effect on stock volatility [3], our study includes stock return kurtosis and skewness to test the potential predictive power of risk factors in investment using different time windows.

III. Data and Methodology

A. Risk Topic Modeling

To quantify the risk disclosure in financial reports, we implement Sentence Latent Dirichlet Allocation (Sent-LDA) model proposed by [22]. Latent Dirichlet Allocation (LDA) is a probabilistic model for a collection of textual data in which each document is represented by a mixture of topics [23]. However, its bag-of-word assumption is not appropriate on risk topics in financial disclosures. Sent-LDA model extends original LDA model by assigning the same topic on words from the same sentence instead of considering words separately, which is able to get more accurate results. Table I summarizes some of the notations used in this study.

TABLE I: Notations

N	Total number of documents.
S_d	Total number of sentences in document $d \in \{1, 2, ...N\}$.
K	Total number of topics. We fix $K = 30$ motivated by [3].
$\mathcal{V}$	Set of vocabulary constructed from the set of documents in the study.
V	$= \|\mathcal{V}\|$, total number of words in vocabulary.
W_d	Total number of words in document d.
α	Parameter of Dirichlet distribution estimated from Sent-LDA model.
β	$V \times K$ probability matrix estimated from Sent-LDA model.

The Sent-LDA model uses two hidden latent variable θ and z with following assumptions:

- Each document d is represented by a mixture of topics, and the proportion of topics is represented by a vector $\theta_d \sim Dirichlet(\alpha)$.

- z_d is a S_d by K matrix represents probability distribution of sentences in document d over topics. For each sentence s in document d satisfies $z_{d,s} \sim Multinomial(\theta_d)$.

The Sent-LDA model involves solving an inferential problem and thus the computation of the posterior distribution $p(\theta, z|w, \alpha, \beta)$ of two hidden variables θ and z. Due to the intractable computation on the distribution, there are two algorithms usually used for topic modeling: sampling-based method and variational inference [24]. Although sampling-based method is much easier to implement, variational inference is an optimized method which generates more accurate results [3]. In this study, we use variational inference to approximate the intractable distribution.

The idea of convexity-based variational inference is to obtain a lower bound for log-likelihood of the observed data from a family of approximated distributions, which is characterized by the following variational distribution with two free variational parameters γ and ϕ:

$$q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \prod_{k=1}^{K} q(z_k|\phi_k)$$

Following the algorithm in [23], the optimal values of variational parameters γ and ϕ are found by minimizing the Kullback-Leibler divergence between the variational distribution and true posterior distribution. An iterative method is used in the optimization process to update the value of γ and ϕ by setting the first derivatives of KL divergence to zero. In each iteration, the derivation of variational EM algorithm is used to find the optimizing values of γ and ϕ (E-Step) and maximize the lower bound of log-likelihood with respect to α and β (M-Step). The following update equations for γ and ϕ are obtained from variational inference used in E-step, and the update equation for β is used in M-Step:

$$\phi_{d,s,k} = \left(\prod_{w=1}^{W_{d,s}} \beta_{w,k} \right) exp\left(\psi(\gamma_{d,k}) - \psi\left(\sum_{i=1}^{K} \gamma_{d,i} \right) \right)$$

$$\gamma_{d,k} = \alpha + \sum_{s=1}^{S_d} \phi_{d,s}$$

$$\beta_{k,v} = \sum_{d=1}^{N} \sum_{s=1}^{S_d} \sum_{w=1}^{W_{d,s}} \sum_{k=1}^{K} \phi_{d,s,w,k} w_{d,s,w}^{v}$$

where $\phi_{d,s,k}$ is the probability that sentence s is generated by topic k in document d, $\gamma_{d,k}$ is the kth component of posterior Dirichlet distribution in document d, $\beta_{k,v}$ is the probability that vth word is generated by topic k, $w_{d,s,w}^{v} = 1$ if the word in the sentence $Word_w$ is the same with the work in the vocabulary $Word_v$ and $w_{d,s,w}^{v} = 0$ otherwise, and $\psi(x)$ is the first derivative of $log\Gamma(x)$. The detailed iterative process of Sent-LDA model for parameter estimation is shown in Algorithm 1.

The β matrix is used to calculate the probability distribution of topics on one sentence as following:

$$p(s|z) = \prod_{w=1}^{W_{d,s}} \beta_{z,v_w}$$

where β_{z,v_w} is the probability of wth word in the sentence about topic z.

We choose the topic with maximum probability assigned to the sentence, and we can then calculate the frequency of topic z appears in one document d as

$$d_z = \sum_{s=1}^{S_d} Sentence_{s,z}$$

where $Sentence_{s,z} = 1$ when topic z is assigned on the sentence, and $Sentence_{s,z} = 0$, otherwise. d_z is used as risk disclosure variables in the multi-factor models in section IV.

Algorithm 1 Variational inference algorithm for Sent-LDA

```
Initialization
α = 1.
β_{k,v} = 1/V + Rand, then β is normalized to Σ β_{k,·} = 1, where Rand is random number between 0 and 1, k ∈ {1, 2, ...K}, v ∈ {1, 2, ...V}.
do
    for d = 1 to N do (E-Step)
        Initialization: φ = 1, γ = α + S_d/K
        for s = 1 to S_d do
            for k = 1 to K do
                Update φ
            end for
        end for
        Update γ
        Compute lower bound of log likelihood
        L_d(γ, φ; α, β)
    end for
    Update β
while Δ(L) < 10^{-5} ≈ 0
where Δ(L) = Σ_{d=1}^{N} L_d^{new}(γ, φ; α, β) − Σ_{d=1}^{N} L_d^{old}(γ, φ; α, β)
```

B. Topic Matching Process

Our second objective is to match new risk factors with the existing risk types based on word probability distributions. A possible approach to address this is topic detection and tracking (TDT), previously used to track the events in news [25]. Another possible approach considers topic models on different time slices and parameters in time t are associated with prior parameters in $t-1$ [26]. However, this approach is restricted to the same number of topics and it is a sequential process without parallel processing. Recent studies have improved the topic matching by involving similarity measure between two word probability distributions which is equivalent to calculating the similarity between two topics from separate topic models [27]. There are multiple similarity measures in text categorization field, such as Kullback-Leibler (KL) divergence, cosine similarity and dices coefficient [27], [28].

In this study, we apply cosine similarity measure on normalized word probability distributions among N words with highest probability from Sent-LDA model to get the similarity

matrix. If two topics relate to the same risk factor, they generally have similar words with high probability and the value of cosine similarity ($sim(z_i^{New}, z_j^{Old})$) should be closer to 1. For topic matching process, previous studies have used several methods, such as support vector machine (SVM) [29] and neural network [30]. Because our goal is to cluster 30 topics into pairs, we adopt hierarchical clustering method [31] to match topics based on distance matrix converted from the similarity matrix. First of all, we build a distance matrix between existing topics (z^{Old}) and new topics (z^{New}) by using $dist(z_i^{New}, z_j^{Old}) = 1 - sim(z_i^{New}, z_j^{Old})$ to represent distance between z_i^{New} and z_j^{Old}. We define $dist(z_i^{New}, z_j^{New}) = dist(z_i^{Old}, z_j^{Old}) = 1$. Then, we apply hierarchical clustering method on the distance matrix to get topic pairs. Because the distances between two existing topics and two new topics are assigned to maximum distance, each topic pair contain one new topic and one existing topic. If one topic contains no common word with all existing topics, it will be assigned to single cluster. The detailed process of topic matching algorithm are shown in Algorithm 2.

Algorithm 2 Topic Matching Process

1: Get K topics ($z_1^{Old}, z_2^{Old}, ..., z_K^{Old}$) using Sent-LDA model on existing dataset.
2: Get K' topics ($z_1^{New}, z_2^{New}, ..., z_{K'}^{New}$) using Sent-LDA model on new dataset.
3: Normalize probability of N words with highest probability $p(\mathbf{w}|z_i^{New})$ from z_i^{New}.
4: Get the probability of the same N words from z_j^{Old} as $p(\mathbf{w}|z_j^{Old})$. If there exists a word $w \in z_i^{New}, w \notin z_j^{Old}$, $p(w|z_j^{Old}) = 0$. Then, we normalize $p(\mathbf{w}|z_j^{Old})$.
5: Calculate cosine similarity $sim(z_i^{New}, z_j^{Old}) = \frac{z_i^{New} \cdot z_j^{Old}}{||z_i^{New}||||z_j^{Old}||}$ for each z_i^{New} and z_j^{Old}.
6: Calculate distance matrix using:
   - $dist(z_i^{New}, z_i^{New} j) = dist(z_i^{Old}, z_j^{Old}) = 1$
   - $dist_(z_i^{New}, z_j^{Old}) = 1 - sim(z_i^{New}, z_j^{Old})$
7: Run hierarchical clustering on distance matrix and get topic pairs.

C. Data Collection

We collect whole textual part of Section 1A in annual financial filings (10-K) submitted to the SEC from 2011 to 2015 including 22,991 filings (See Table II). The textual data contain all characters in risk disclosures. In Section 1A, filers usually state multiple risk factors, each of which often includes one sentence as subtitle followed by several paragraphs. Following [3], we only use the subtitles as the input data in the Sent-LDA model. We believe that the entire text and paragraphs in Section 1A include too detailed information which may reduce the accuracy of the topic extraction procedure.

Then, we extract subtitles from the textual data. In [3], html style tag is used for data extraction; hence, subtitles are specified by some special tags (Bold or Italic). However, our data do not contain any tags and we can only use line break[1] as identifier for a subtitle. Using random sampling validation of these extracted subtitles, we find that most subtitles are single sentence separating from other text paragraphs. Thus, we apply the following rules to extract subtitles in Section 1A:

- Separate into paragraphs by line break character ("\n").
- Check number of period (".") in each paragraph. If the paragraph includes no period or only one period, it is marked as *subtitle*.
- For each *subtitle*, irrelevant English words are excluded[2]

As a result, the final dataset includes multiple subtitles for each risk disclosure, and each subtitle is represented by a vector of words. However, our rules are not able to filter non-subtitle paragraphs with single sentence in body part. Compared with dataset in [3], our dataset has more sentences in average on each document.

For stock prices data collection, we use the companys Central Index Key (CIK) as the identifier to download daily stock close price data from COMPUSTAT database. We only keep financial filings with complete stock records and subtitle data for further analysis. Also, if one company (CIK) has been matched to multiple stock tickers (e.g. GOOG and GOOGL), we only keep the one with the smaller number of letters, e.g., we use GOOG. Thus, we can ensure one row stock data for each financial filing. As a result, there are total of 16,110 filings from 4,919 unique companies in our analysis.

TABLE II: Data Summary

Year	2011	2012	2013	2014	2015	Total
Total Filings	1,319	5,466	5,460	5,423	5,323	22,991
With Ticker	1,070	3,736	3,759	3,937	3,608	16,110

IV. Empirical Study

A. Topic Matching

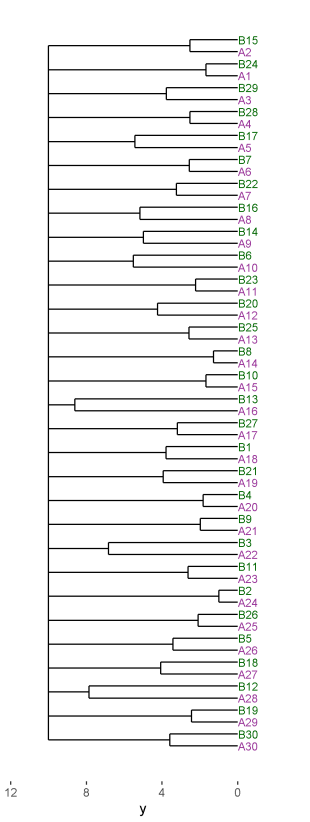
In this section, we implement Sent-LDA model and hierarchical clustering to extract topics from risk disclosures and match topics from different models. We use data from 2011 to 2014 as training dataset and data from 2015 as testing dataset for empirical analysis. First, we apply Sent-LDA model to extract 30 topics from training dataset as *existing topics* (z_i^{Old} where $i \in \{1, 2, ..., 30\}$), and 30 topics from testing dataset as *new topics* (z_j^{New} where $j \in \{1, 2, ..., 30\}$). Topic z is represented by $\beta_{\cdot,z}$, a vector of probability distribution among words, which is used to calculate cosine similarity between two topics. We observe that words with higher probability are much more meaningful for distinguishing a topic. Thus, we only use N words with highest probability to measure topic similarity. After calculating CDF on descending sorted word probability (see Figure 1 as an example), we find that the cumulative probability of 400 high-probability words already

[1]The textual data includes "\n" as line break character.

[2]We delete numbers and non-English words. Also, we delete stop words such as "the", "a", "this", etc.. In additional, we excluded irrelevant phrases such as "content table".

TABLE III: 30 Risk Factors from Sent-LDA Model

A[a]	B[b]	Match	Risk Factor [c]	Feature Words
1	24	Y	Regulation Change	regulation, change, operation, environment
2	15	Y	Tax Risks	tax, distribution, income, federal
3	29	Y	Corporate Management[d]	company, requirement, governance, regulation
4	28	Y	Property	gas, oil, natural, price
5	17	Y	Financial Condition	asset, goodwill, impairment, intangible
6	7	Y	Funding	capital, additional, credit, business
7	22	Y	Stock	stock, share, preferred, future
8	16	Y	Accounting	financial, accounting, reporting, internal
9	14	Y	Property	properties, tenant, condition, estate
10	6	Y	Business Condition[d]	business, customer, revenue, demand
11	23	Y	Product Introduction	product, approval, candidate, clinical
12	20	Y	Debt Risks	debt, indebtedness, covenant, credit
13	25	Y	Cost/Supply	cost, material, supplier, customer
14	8	Y	Acquisition	acquisition, business, risk, future
15	10	Y	Human Resources	personnel, retain, management, attract
16	12		Stock Volatile	stock, dividend, pay, future
17	13		Stock Volatile	stock, trading, share, market
18	1	Y	Financial Condition	loss, company, revenue, history
19	21	Y	Product Defects Lawsuits	product, litigation, liability, claim
20	4	Y	Property Intellectual	property, intellectual, protect, business
21	9	Y	Macroeconomic Cyclical Industry	economic, market, condition, industry
22	30		Customer	customer, contract, revenue, client
23	11	Y	Infrastructure	information, system, operation, security
24	2	Y	Regulation Changes	provision, takeover, control, change
25	26	Y	Competition	competition, market, service, revenue
26	5	Y	Tax Risks	tax, change, financial, rate
27	18	Y	Investment	investment, loan, mortgage, interest
28	27	Y	Stock Volatile	stock, market, fluctuate, volatile
29	19	Y	Stockholder Interest	stockholder, director, interest, control
30	3	Y	Product Defects Lawsuits	product, liability, claim, insurance



[a] Sent-LDA model topic using data from 2011 to 2014.
[b] Sent-LDA model topic using data from 2015.
[c] Risk factors are from [3].
[d] Newly added risk factors.

exceeds 90% which is sufficient to represent the feature of one topic. We further test different $N = \{50, 100, 200, 300, 400\}$ and find stable result when $N > 300$. We choose to use normalized probability among 300 words with highest probability to measure the cosine similarity matrix.

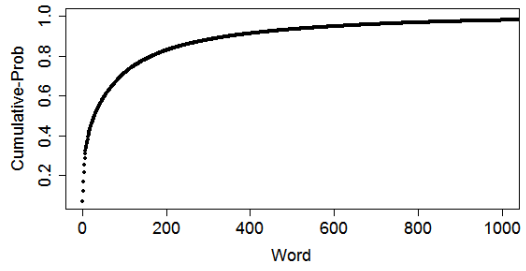


Fig. 1: CDF of Sorted $p(\mathbf{w}|z)$ on 1000 Words on Regulation Risk from Training Dataset

We first extract and normalize probabilities of 300 words with highest probability from z_i^{New} and the probability of the same 300 words from z_j^{Old} is extracted and normalized. We then apply the hierarchical clustering method to match z_i^{New} with z_j^{Old}. We follow Algorithm 2 in previous section to get 60×60 distance matrices among z_i^{New} and z_j^{Old}.

After running the topic matching algorithm, we are able to get matched topic pairs. We include a sample result in Table III comparing risk factors from 2011-2014 dataset and factors from 2015 dataset. We find that 27 out of 30 topics from 2015 can be correctly matched to existing topics. We assign the existing risk factors in Bao and Datta's study [3] for 28 topics of group A based on word cloud results. Topic A3 has words related to company's management, and topic A6's words are more related to business. The existing risk types do not include related topics. We assign new labels on these 2 topics as "Corporate Management" and "Business Condition". Also, our results only matches to part of the existing risk factors. There are 4 risk factors not included in our results, i.e. catastrophes input price risk, international risk, credit risk, downstream risk. The time period of data in the previous study is from 2006 to 2010 which covers the financial crisis in 2008. It might be the reason that we do not see the 4 risk factors. Therefore, it is reasonable to assume that the difference of time period potentially changes the risk factors in disclosure. However, our results show that most risk factors are common

among all years.

To test the robustness of our topic-matching method, we apply the same algorithm on 4 pairs of Sent-LDA model using a rolling window from 2011 to 2015 (See Table IV). While comparing two single-year data, we could successfully match 24 out of 30 topics (80%). The matching accuracy is increased when including more data in training model. When more than two years of data is used (2011-2012, 2011-2013, 2011-2014), we find more than 90% of topics matched correctly.

TABLE IV: Stability of the Topic Matching Process

Data	2011 vs. 2012	2011-2012 vs. 2013	2011-2013 vs. 2014	2011-2014 vs. 2015
Matched	80% (24/30)	90% (27/30)	93% (28/30)	90% (27/30)

B. Risk Factor Model

We use 30 risk factors from Sent-LDA model as input variables and fit a multi-factor model to investigate the potential effects of different risk factors on stock returns. We apply factor model to predict 3 return variables, i.e. stock return volatility (standard deviation), skewness and kurtosis. Moreover, to investigate the time effect of risk factors, we implement the factor model for the same variables on different time window including 1-week, 2week, 3-week, 1-month and 2-month ($T = \{7, 14, 21, 30, 60\}$). The details of variables and models are shown in Table V.

TABLE V: Empirical Factor Models

$Model_{\sigma}^{T}$:	$\sigma_d^{\text{After}} = \sigma_d^{\text{Before}} + \sum_{j=1}^{30} d_j$
$Model_{\gamma}^{T}$:	$\gamma_d^{\text{After}} = \gamma_d^{\text{Before}} + \sum_{j=1}^{30} d_j$
$Model_{\kappa}^{T}$:	$\kappa_d^{\text{After}} = \kappa_d^{\text{Before}} + \sum_{j=1}^{30} d_j$
σ_d^{After}, σ_d^{Before}	Stock return volatility after/before filing date.
γ_d^{After}, γ_d^{Before}	Stock return skewness after/before filing date.
κ_d^{After}, κ_d^{Before}	Stock return kurtosis after/before filing date.
d_j	The frequency of risk factor j appears in the document d.
T	Each model uses stock return T days before and after filing date to calculate σ, γ, κ.

We apply the model on collected data from 2011 to 2015 with 30 risk factors. Due to the space limitation, we only include detailed results for one long-term model and one short-term model ($Model^{T=60}$ and $Model^{T=14}$) in Table VI). All the models on return volatility are significant with higher R-squared except 1-week model, which indicates significant long-term impact of risk factors on stock return volatility. The linear models on skewness and kurtosis have low R-square value which indicates lower linear relationship between risk factors and the two variables. The significance of single risk factor in different models provide more information on the various impact on stock return. Based on the sign of coefficient, we separate the significant variables into two groups which positively/negatively affects post-disclosure stock return.

In the volatility model, acquisition risk, debt risk, accounting risk are the 3 common risk factors affect both short-term and long-term stock volatility. Financial condition risk has impact on short-term volatility in 1-week and 3-week models, and customer risk and funding risk only affect on 1-month model. The above risk factors have positive impacts on stock volatility, which implies that disclosing those risks increases the uncertainty on stock returns. Stockholder risk, stock volatile risk, cost/supply risk, and product introduction risk are common risk factors negatively affecting volatility. Down-stream risk and property risk only impact short-term volatility. The disclosure of such risks tend to reduce the stock volatility by revealing more information to investors. And risks related to company's financial condition are more likely to increase the uncertainty of company's future performance.

Skewness and kurtosis are higher moments of stock returns which also attracted researchers' attention. Firms with lower stock return skewness and kurtosis are more likely getting higher average returns and lower risk [32]. The impact of risk factors on skewness and kurtosis of stock return can help investors evaluate future stock performance. The results show that investment risk and stock volatile risk are the common factors which have positive impact on skewness. Acquisition risk, funding risk and tax risk are positively associated with long-term skewness while macroeconomic risk affects only short-term skewness. Human resource risk impacts the skewness negatively. Catastrophes risk and regulation risks have short-term negative impacts and product related risks have long-term negative impacts. Accounting risk and customer risks only have impacts on 3-week model. Models on kurtosis have more significant factors. Funding risk, acquisition risk, debt risk, human resources risk and accounting risk are common factors with positive effects on kurtosis, and financial condition risk only has positive impact on short-term model. Stock related risks, property risk, down-stream risk and management risk have negative effects on kurtosis.

Comparing volatility, skewness and kurtosis models, acquisition risk has positive effect on all three models. It indicates that potential acquisition activity significantly changes investors' perception on the performance of company. Stock related factors, including stock volatile risk, stockholder risk, shareholder interest risk, downstream risk and cost/supply risk are negatively associated with volatility and kurtosis. Disclosing such risks send information to investors and reduce their perception on future stock risk. According to the results, the significant risk factors have predictive power on company's stock risk. Impacts from different risk factors provide more detailed information on both positive and negative changes of post-disclosure stock risk instead of single direction prediction.

TABLE VI: Estimation Results on $Model^{T=60}$ and $Model^{T=14}$

	Volatility model (σ)		**Skewness model (γ)**		**Kurtosis model (κ)**	
Variable	$T=60$ Coef	$T=14$ Coef	$T=60$ Coef	$T=14$ Coef	$T=60$ Coef	$T=14$ Coef
Investment	-0.0008	0.0004	0.0306**	0.0149*	0.0120	-0.0019
Acquisition	0.0010**	0.0011**	0.0226***	0.0057	0.0554*	0.0245***
Funding	0.0002	0.0001	0.0050*	-0.0015	0.0329***	0.0123***
Stockholder	-0.0007	-0.0008*	0.0018	0.0037	-0.1257***	-0.0200***
Tax	-0.0003	0.0003	0.0206***	0.0037	0.0259	0.0029
Product	-0.0001	-0.0005	-0.0007	0.0005	-0.0001	-0.0072
Debt	0.0017***	0.0013***	0.0024	-0.0015	0.1250***	0.0257***
Stock Volatile	-0.0010***	-0.0011**	-0.0025	0.0021	-0.1060***	-0.0246***
Human Resource	-0.0001	-0.0001	-0.0056*	-0.0051***	0.0140	0.0050*
Cost/Supply	-0.0014**	-0.0012	0.0021	0.0028	-0.1255***	-0.0527***
Accounting	0.0026***	0.0019 ***	-0.0003	-0.0035	0.3038***	0.0675***
Competition	0.0003	0.0001	-0.0076	0.0019	-0.0361	-0.0060
Downstream	-0.0004	-0.0010**	-0.0021	-0.0018	-0.0899***	-0.0219***
Investment	-0.0001	-0.0002	-0.0024	0.0031	0.0535**	0.0054
Infrastructure	-0.0004	-0.0003	-0.0041	-0.0024	-0.0148	-0.0032
Intellectual Property	-0.0002	-0.0005	-0.0059	0.0063	-0.0563**	-0.0199***
Properties	-0.0002	-0.0002	-0.0044	-0.0078	-0.0465	-0.0136*
Tax	0.0002	0.0001	-0.0076	0.0004	0.0100	0.0059
Corporate Management	-0.0003	0.0001	-0.0006	-0.0055	-0.0573***	-0.0106**
Macroeconomics	-0.0002	-0.0001	-0.0019	0.0031*	-0.0014	-0.0034
Product Introduction	-0.0011***	-0.0008*	-0.0140*	-0.0103**	-0.0954***	-0.0227***
Regulation Change	-0.0001	0.0001	0.0016	-0.0006	-0.0113	-0.0010
International	-0.0005	-0.0002	-0.0084	-0.0058	-0.0787***	-0.0017
Catastrophes	-0.0003	-0.0002	-0.0085	-0.0065*	-0.0400*	0.0023
Shareholder Interest	-0.0007***	-0.0010***	-0.0060	0.0035	-0.0578***	-0.0192***
Stock Volatile	0.0001	0.0007	0.0248***	0.0077*	-0.0108	0.0030
Financial Condition	0.0001	0.0004	-0.0042	0.0032	0.0107	0.0126**
Regulation Change	0.0002	0.0002	-0.0047	-0.0068**	0.0265	-0.0104**
Customer	0.0002	0.0005	-0.0100	-0.0076	-0.0348	-0.0009
Product Lawsuit	-0.0001	-0.0003	-0.0049	-0.0026	0.0014	-0.0041
Intercept	0.0138***	0.0111***	0.0328*	0.0488***	2.0271***	-0.2821***
σ^{Before}	0.7152***	0.7730***				
γ^{Before}			-0.0142*	-0.0171**		
κ^{Before}					0.3116***	0.2094***
R^2_{Adj}	0.5484	0.5190	0.0028	0.0023	0.1282	0.0729
p-value	0.0000	0.0000	0.0000	0.0000	0.0002	0.000
F-Stat	631.60	561.40	2.42	2.16	76.22	40.18

$^{***}p < 0.01$ $^{**}p < 0.05$ $^{*}p < 0.10$

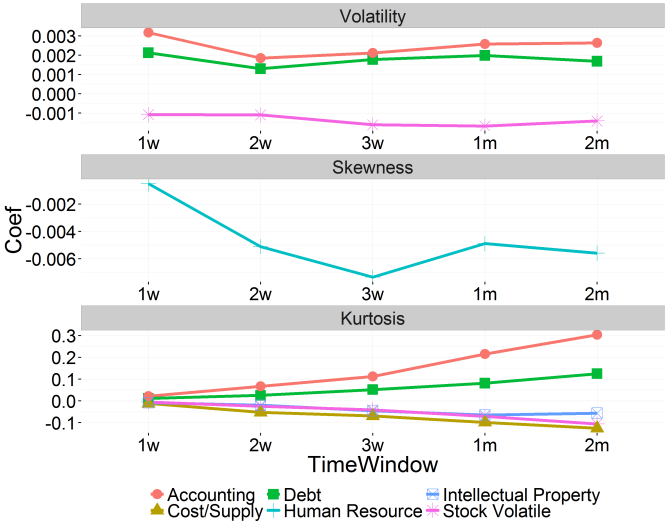


Fig. 2: Sensitivity of Stock Return to Risk Factors Over Time

Furthermore, to investigate the impact progression of risk factors on short-term and long-term stock returns, we compare the coefficients of the common risk factors. Figure 2 shows the coefficient changes of accounting, debt, stock volatile risks in volatility model, human resource risk on skewness model and accounting, debt, intellectual property, cost/supply and stock volatile risks on kurtosis model. In volatility model, the two positively affected factors have highest impacts in 1-week model and the impact decreases to a lower level as the time window increases. The stock volatile risk has less negative impact on the short-term volatility and the impact increases in long-term models. Skewness also receives greatest impact from human resource factor in 1-week model, and the lowest impact is in 3-week model. The progression in kurtosis models is more significant. Both positive risk factors (accounting and debt) and negative risk factors (intellectual property risk, cost/supply risk and stock volatile risk) have less impacts on 1-week model and the effects increased on

long-term kurtosis. The results show that most of the risk factors have more effects on short-term stock volatility and skewness, and more impact on long-term stock return kurtosis. It is possibly caused by the feature of kurtosis which measures the "tail" of stock return. With the continuous impact from risk factors, the change of stock return distribution will become more significant in long term.

V. Conclusion

In this paper, we investigate the effect of risk factors on stock returns using financial risk disclosure textual data. We apply an unsupervised topic modeling method, Sent-LDA, on risk disclosures from public companies' annual reports from 2011 to 2015 to extract risk factors. We propose a novel algorithm to match topics from new dataset to existed risk factors. We test the algorithm on different datasets, and more than 90% topics can be correctly labeled to the corresponding risk factors.

Then, we include 30 extracted risk factors in multi-factor models to investigate their impact on stock returns. We test the impact on stock return volatility, skewness and kurtosis which represent stock returns risk characteristics. Our results show that some risk factors, such as acquisition risk, are positively associated to stock return risk. Disclosing such risks tends to increase investors risk perception and the uncertainty in the market. On the other hand, more risk factors, such as stock related risks, are negatively associated with three variables. It indicates that the disclosure of these risk factors indeed helps investors on risk evaluation and reduces the uncertainty of the market. Disclosed risk factors also provide various impacts on stock returns with different time window. Moreover, we observe the progression of the common risk factors and find that they have significant impact on short-term volatility and skewness, and long-term kurtosis. Overall, we find that the risk factors from the risk disclosures have significant correlation with stock return volatility, and the risk disclosures provide better information to investors and they do change investors' risk perception overtime.

In the future, we plan to extend the topic matching algorithm by adding filter conditions to distinguish emerging risk factors. The extended algorithm will be able to map traditional risk factors and notify the user with new risks which can improve the accuracy of topic matching. Furthermore, we plan to apply risk factors extracted from Sent-LDA model to construct and re-balance stock portfolios for better and robust alphas.

References

[1] B. D. Cadman, T. O. Rusticus, and J. Sunder, "Stock option grant vesting terms: economic and financial reporting determinants," *Review of Accounting Studies*, vol. 18, no. 4, pp. 1159–1190, 2013.

[2] R. A. DeFusco, D. W. McLeavey, J. E. Pinto, M. J. Anson, and D. E. Runkle, *Quantitative investment analysis*. John Wiley & Sons, 2015.

[3] Y. Bao and A. Datta, "Simultaneously discovering and quantifying risk types from textual risk disclosures," *Management Science*, vol. 60, no. 6, pp. 1371–1391, 2014.

[4] J. L. Campbell, H. Chen, D. S. Dhaliwal, H.-m. Lu, and L. B. Steele, "The information content of mandatory risk factor disclosures in corporate filings," *Review of Accounting Studies*, vol. 19, no. 1, pp. 396–455, 2014.

[5] SEC, "Securities and exchange commission final rule, release no. 338591 (fr-75)." 2005. [Online]. Available: http://www.sec.gov/rules/final/33-8591.pdf

[6] SEC, "Annual report pursuant to section 13 or 15(d) of the securities exchange act of 1934, general instructions," 2010. [Online]. Available: http://www.sec.gov/about/forms/form10-k.pdf

[7] F. Li, "Do stock market investors understand the risk sentiment of corporate annual reports?" *Available at SSRN 898181*, 2006.

[8] R. Feldman, S. Govindaraj, J. Livnat, and B. Segal, "Managements tone change, post earnings announcement drift and accruals," *Review of Accounting Studies*, vol. 15, no. 4, pp. 915–953, 2010.

[9] T. Kravet and V. Muslu, "Textual risk disclosures and investors risk perceptions," *Review of Accounting Studies*, vol. 18, no. 4, pp. 1088–1122, 2013.

[10] F. Li, "Textual analysis of corporate disclosures: A survey of the literature," *Journal of accounting literature*, vol. 29, p. 143, 2010.

[11] M.-A. Mittermayer, "Forecasting intraday stock price trends with text mining techniques," in *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*. IEEE, 2004, pp. 10–pp.

[12] E. Henry, "Are investors influenced by how earnings press releases are written?" *Journal of Business Communication*, vol. 45, no. 4, pp. 363–407, 2008.

[13] F. Li, "The information content of forward-looking statements in corporate filingsa naïve bayesian machine learning approach," *Journal of Accounting Research*, vol. 48, no. 5, pp. 1049–1102, 2010.

[14] C. M. Callahan and R. Smith, "Firm performance and management's discussion and analysis disclosures: An industry approach," *Available at SSRN 588062*, 2004.

[15] J. Boudoukh, R. Feldman, S. Kogan, and M. Richardson, "Which news moves stock prices? a textual analysis," National Bureau of Economic Research, Tech. Rep., 2013.

[16] F. Li, "Annual report readability, current earnings, and earnings persistence," *Journal of Accounting and economics*, vol. 45, no. 2, pp. 221–247, 2008.

[17] T. Loughran and B. McDonald, "Measuring readability in financial text," *SSRN eLibrary*, 2010.

[18] Y. Mirakur, "Risk disclosure in SEC corporate filings," *Wharton Research Scholars Journal*, p. 85, 2011.

[19] K. K. Nelson and A. C. Pritchard, "Litigation risk and voluntary disclosure: The use of meaningful cautionary language," in *2nd Annual Conference on Empirical Legal Studies Paper*, 2007.

[20] K.-W. Huang and Z. Li, "A multilabel text classification algorithm for labeling risk factors in sec form 10-k," *ACM Transactions on Management Information Systems (TMIS)*, vol. 2, no. 3, p. 18, 2011.

[21] S. Aral, P. G. Ipeirotis, and S. J. Taylor, "Content and context: Identifying the impact of qualitative information on consumer choice," *Available at SSRN 1784376*, 2011.

[22] Y. Bao and A. Datta, "Summarization of corporate risk factor disclosure through topic modeling," 2012.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[24] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[25] J. Allan, *Topic detection and tracking: event-based information organization*. Springer Science & Business Media, 2012, vol. 12.

[26] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 113–120.

[27] A. Niekler and P. Jähnichen, "Matching results of latent dirichlet allocation for text," in *Proceedings of ICCM*, 2012, pp. 317–322.

[28] B. Bigi, "Using kullback-leibler distance for text categorization," in *European Conference on Information Retrieval*. Springer, 2003, pp. 305–319.

[29] D. Quercia, H. Askham, and J. Crowcroft, "Tweetlda: supervised topic classification and link prediction in twitter," in *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 2012, pp. 247–250.

[30] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," *arXiv preprint arXiv:1412.1058*, 2014.

[31] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[32] J. Conrad, R. F. Dittmar, and E. Ghysels, "Ex ante skewness and expected stock returns," *The Journal of Finance*, vol. 68, no. 1, pp. 85–124, 2013.