Comparative Text Analytics via Topic Modeling in Banking

Yu Chen*, Rhaad M. Rabbani*, Aparna Gupta†, and Mohammed J. Zaki*
*Computer Science Department, Rensselaer Polytechnic Institute, Troy, NY 12180.
†Lally School of Management, Rensselaer Polytechnic Institute, Troy, NY 12180

***Abstract*—In this paper, we compare and evaluate multiple topic modeling approaches and their effectiveness in analyzing a large set of SEC filings by US public banks. More specifically, we apply four major topic modeling methods to a corpus of 8-K and 10-K filings, from the years 2005–2016, of 578 bank holding companies. These methods include Principal Component Analysis, Non-negative Matrix Factorization, Latent Dirichlet Allocation and KATE, a novel k-competitive autoencoder for text documents. Separately for 8-K and 10-K, the usefulness and effectiveness of these methods is evaluated by comparing their performances on two classification tasks: (i) predicting which section each document corresponds to, where we consider each section within an 8-K or 10-K filing as an individual document, and (ii) detecting text from a bank's year of failure, a task for which we use bank failure data from the 2008 financial crisis. In addition, we qualitatively compare the topics discovered by the different methods. We conclude that topic modeling can be an effective tool in financial decision making and risk management.**

I. Introduction

Over the past several years, research in finance and banking has acknowledged and started to utilize large amounts of textual data available through reports, regulatory filings, print news media, social media platforms, chat rooms, and discussion boards. Taking advantage of these textual data for important financial and risk insights is a valuable pursuit, especially in a manner that these insights can complement those obtained from quantitative data. Finance and banking specific text analytics research challenges can benefit from expertise of both finance researchers and computer science experts. In this paper, we further this objective by developing comparative analytics for a range of classical and novel topic analysis techniques for a large set of banks using their SEC filings.

In finance literature, a large fraction of text analytics work has been devoted to predicting equity price and other market variable movement. For example, Alfano et al. [1], Antweiler and Frank [2] and Wuthrich et al. [3] use news articles to predict the stock market or FOREX market. Serrano and Iglesias [4], Nguyen et al. [5] and Ranco et al. [6] focus on analyzing social media text from platforms such as Twitter and Yahoo Finance message board, for implementing market prediction. Besides market prediction, text mining has also been utilized for measuring financial conditions, such as credit rating prediction [7], [8], bank distress prediction [9], [10] and systemic risk measurement [11].

The application of text mining in finance usually follows certain steps. Feature extraction methods determine how researchers collect useful qualitative information from textual data. The most popular method used for feature extraction is the bag-of-words approach [1], [5]. This technique breaks the text into word-level units, and treats these units as features, while ignoring the order and co-occurrence of words [12]. Schumaker et al. [13] apply a noun-phrase technique, by identifying words as noun part-of-speech (PoS) using a lexicon and then apply syntactic rules to detect noun phrases around that noun to extract features. Furthermore, Vu et al. [14] implement a named entity recognition technique on tweets and improve the feature extraction results. After feature extraction, the next step is usually feature selection, followed by application of a classification method to capture the required signal from the text. Various machine learning algorithms have been applied to analyze the features extracted [5], [15], [16].

A bag-of-words (word frequency) based textual analysis runs the risk of getting too conditioned by the relevance and currentness of the word dictionary chosen. In contrast, *in this paper our objective is to compare and evaluate topic modeling approaches and their effectiveness in analyzing a large set of SEC filings made by US public banks*. We apply the methods to a corpus of 8-K and 10-K filings of 578 bank holding companies. The effectiveness of the methods is evaluated by (i) their ability to identify the document sections the text was extracted from, and (ii) their ability to detect text from a bank's year of failure. We find that some of the novel methods perform at high level of accuracy at both tasks even without needing to associate any sentiments with the topics [17].

II. Methods

We employ several unsupervised learning algorithms to learn topics from 8-K and 10-K filings, which are the most important mandatory filings required of all public firms. In this paper, as our focus is on banks, the 2005-2016 period offers a natural experiment when many banks failed during the global financial crisis. In natural language processing, a topic is typically modeled as a group of words, where each word has an associated membership weight. From a geometric viewpoint, a topic is a vector in word space, where every dimension corresponds to a different word, and the components of the vector are the membership weights of the respective words. Similarly, we model a document (a bag of words) as a group of topics, where each topic has a membership weight. A document is a point in topic space.

Among the algorithms we study, Non-negative Matrix Factorization (NMF) [18] and Latent Dirichlet Allocation (LDA) [19] are known well for their topic modeling capabilities. We also allow negative membership weights via the use of Principal Component Analysis (PCA) [20]. Finally, we use KATE [21], a novel auto-encoder based approach, as a topic modeling method. For a given topic space dimensionality, each of the above methods generates a topic model, which provides a topic space representation for any input document. These unsupervised learning algorithms used in our study are detailed below.

A. LDA

Probabilistic topic models, such as probabilistic Latent Semantic Analysis (pLSA) [22] and Latent Dirichlet Allocation (LDA) [19], typically model a document as a mixture of topics and a topic as a mixture of words. Many variants have been proposed to tackle different problems emerging in

topic modeling. Most topic models take a topic as a bag of words. However, phrases can sometimes help discover more interpretable topics as they are more informative than the sum of their individual components. For example, "white house" as a phrase carries a special meaning under the "politics" topic instead of "a house which is white" under the "real estate" topic. Many models [23]–[25] have been proposed to relax the bag-of-words assumption. In general, individual documents usually focus on a few salient topics that typically adopt a narrow range of terms instead of a wide coverage of the vocabulary. Other works have been proposed to introduce sparsity into topic models [26], [27], and to learn dynamic topic models [28], [29].

LDA has been widely used in text analytics. It assumes that documents are constructed as mixtures of latent topics, where each topic is essentially a probability distribution over words. LDA is a graphical model, where the generative process of creating a document is as follows: 1) We randomly sample a topic proportion vector θ which assigns different weights to a set of topics based on a Dirichlet distribution which is parameterized by α and shared by the whole corpus. 2) For each word position in the document, we randomly sample a topic z based on the associated topic proportion vector θ. 3) Given the assigned topic z for the word position, we randomly sample a word from the predefined vocabulary based on a topic-word probability distribution β. The inference process (i.e., estimating θ, β and z) of LDA basically maximizes an approximate posterior. Since LDA itself captures the intuition of document representation, we simply use the topic proportion vector of each document as its representation.

B. PCA

PCA [20] is a multivariate dimensionality reduction technique. In a real world data set, where each data point is of a certain dimension n, a lower dimensional manifold usually accounts for all the data points. Given a target dimensionality k, where k is between 1 and n, PCA discovers the k-dimensional subspace, comprising the k orthogonal vectors called the principal components, that best describe the variance in the data.

Our data points are word space representations of documents. Each principal component, in word space, discovered by PCA represents a distribution of words. We let each principal component represent a topic. We project the word space representation of any document onto this topic space to obtain a topic space representation of the document. In addition to using PCA to extract topics, using it as a dimension reduction method, we also use projections onto the first three components to visualize the clustering of the data points.

C. NMF

Where PCA enforces an orthogonality constraint, NMF enforces a non-negativity constraint. It is a technique for factorizing a data matrix V into non-negatives matrices W and H. The non-negativity constraint causes NMF to learn localized feature representations of the data, where the data matric V is given as the product of W and H, plus a residual matrix U, which represents the error of the factorization. There are a number of different algorithms that are commonly used to achieve a non-negative factorization [18], based on different metrics to measure the error of the factorization.

For our study, we have the document-word matrix as the data matrix V. After performing NMF, W becomes the document-topic matrix and H, the topic-word matrix. The matrix H can be used to find the topic space representation of any document, even a document not used during the factorization.

D. KATE

An autoencoder is a neural network which tries to reconstruct its input at the output layer [30]. An autoencoder consists of an encoder which maps the input $\boldsymbol{x}$ to the hidden layer: $\boldsymbol{z} = g(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$ and a decoder which reconstructs the input as: $\hat{\boldsymbol{x}} = o(\boldsymbol{W}'\boldsymbol{z} + \boldsymbol{c})$; here $\boldsymbol{b}$ and $\boldsymbol{c}$ are bias terms, $\boldsymbol{W}$ and $\boldsymbol{W}'$ are input-to-hidden and hidden-to-output layer weight matrices, and g and o are activation functions. Weight tying (i.e., setting $\boldsymbol{W}' = \boldsymbol{W}^T$) is often used to regularize the model. While vanilla autoencoders even with perfect reconstructions usually only extract trivial representations of the data, more meaningful representations can be obtained by adding appropriate regularization or constraints to the models. Following this line of reasoning, many variants of autoencoders have been proposed recently [31]–[33]. For example, the denoising autoencoder [31] inputs a corrupted version of the data while still trying to reconstruct the original uncorrupted data at the output layer, which forces the model to learn features useful for denoising. As another example, the k-sparse autoencoder [33] explicitly enforces sparsity by only keeping the k highest activities in the feedforward phase.

When examining the features learned by autoencoders, in our recent work [21] we observed that they were not distinct from one another. That is to say, some high frequency words dominate the learned topics. We hypothesized that an autoencoder greedily learns relatively trivial features in order to reconstruct the input as much as possible. To overcome this drawback, we proposed the KATE model in which we force each neuron in the hidden layer to take responsibility for recognizing different patterns within the input data by introducing competition in the training phase. Specifically, in the feedforward phase, after computing the activations $\boldsymbol{z}$ for a given input $\boldsymbol{x}$, the most competitive $\boldsymbol{k}$ neurons are selected as the "winners" while the remaining "losers" are suppressed (i.e., made inactive). However, in order to compensate for the loss of activation from the loser neurons, and to make the competition among neurons more pronounced, the net activations are reallocated among the winner neurons. To respect both the positive and negative patterns captured by the hidden neurons, the aforementioned competition process is done for both positive and negative neurons, respectively.

We apply the above methods in our comparative analytics in the next sections, starting with providing a description of the data and the preprocessing.

III. Datasets

A. Corporate filings

Every plain text and HTML form filed by US corporations with the United States Securities and Exchange Commission (SEC) is available online for public access at SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system [34]. For a selection of 578 bank holding companies (BHCs) corresponding to 52 failed and 526 non-failed banks during the 2008 financial crisis, we retrieve every Form 8-K and Form 10-K filing between the years 2005 and 2016 (inclusive). In

the case of several failed banks, the BHC continued to file beyond the year the bank failed. The list of BHCs and their corresponding failed years was constructed in our previous work [17].

TABLE I
STATISTICS OF 8-K AND 10-K DATASETS

dataset	8-K	10-K
size of training set	44,130	45,160
size of validation set	1,500	1,500
size of test set	12,508	11,616
vocabulary size	4,000	4,000
average length	72	1,187
no. of classes (section types)	8	17

NB: Each data point represents the body of a section.

B. Documents

The retrieved data comprises 50,223 Form 8-K and 4,172 Form 10-K submissions. Each submission contains numbered sections and subsections.

We treat each 8-K or 10-K section label as a class, and each section body within an SEC filing report as an individual document. Thus, the "bag-of-words" representation of each section, i.e., a document, is a labeled data point in the word space defined by a given vocabulary. Thus, even though there are 4,172 10-K submissions, there are over 58K documents or data points (i.e., sections) for the 10-K noted in Table I (the total number of 10-K documents is the sum of the training, testing and validation sets, which equals 58,276 documents/sections).

In the case of a Form 8-K submission, if multiple subsections share the same text, i.e., if multiple subsection headers are listed contiguously followed by text, we match the text to the last subsection header. Finally, we merge all subsections for every top-level section, e.g., all subsections labeled 5.x are merged to form section 5. There are a total of 58,138 final documents for the 8-K dataset, as shown in Table I.

In the case of a Form 10-K submission, we keep sections 1, 1A, 7 and 7A as separate sections, due to their substantive size, but merge subsections under every other numbered section, e.g., 9, 9A and 9B become section 9. For our experiments, we ignore any section text that is included in an exhibit separate from Form 10-K. We also ignore any section that contains fewer than ten words.

C. Preprocessing

To extract sections, we split the entire 8-K or 10-K submissions into paragraphs and use different customized techniques to mark which paragraphs correspond to the beginning of a numbered section (e.g., Item 1A, Item 5.02, etc) or an unnumbered section (e.g., Signatures, Exhibit), or to a page number, or the start or end of a table. We list all the section headers to ensure they are in a sequence, to remove false positives if necessary, and to find instances of section headers that were missed, in which case, we had to further customize our code logic to locate the missing headers. Once we identify the starting paragraph of each section, we extract all the numbered sections.

After extracting each section, we apply a series of natural language processing techniques. These include, in order of application, sentence tokenization, part of speech tagging, removal of stop words and punctuation, and lemmatization (which requires part of speech tags), for which we use the Natural Language Toolkit (NLTK) (www.nltk.org).

For the 8-K and 10-K dataset, respectively, data from 2005 to 2013 becomes our training set, and the remaining data from 2014 to 2016 is used as the test set. We consider the most frequent 4,000 words in the training set as the vocabulary, and consequently represent each document as a bag of words. We randomly select a further 1,500 documents from our training set to be the validation set.

Table I summarizes different statistics for the documents. It lists the training, testing and validation set sizes, the size of the vocabulary, average document length and the number of classes.

D. Comparison of Various Methods

We build topic models from the training and validation datasets for each of the methods in our study, and evaluate and compare the topic models over two classification tasks. The method we compare include: **PCA** [20]: a multivariate dimensionality reduction technique that discovers the k-dimensional subspace that best describes the variance in the data. The principal components that describe the subspace are used as the topic vectors in word space. **NMF** [35]: a matrix factorization technique that factorizes a non-negative matrix V into two non-negative matrices W and H. In our case, W is the document-topic matrix and H is the topic-word matrix. **LDA** [19]: a directed graphical model which models a document as a mixture of topics and a topic as a mixture of words. We used the gensim [36] implementation in our experiments. **KATE** [21]: a k-competitive autoencoder that explicitly enforces competition among the neurons in the hidden layer by selecting the k highest absolute activation neurons as winners, and reallocates the energy from the losers. Our implementation is available at https://github.com/hugochan/KATE.

IV. COMPARISON AND DISCUSSION

We outline our comparative results for the different topic modeling approaches in three steps. In the first step, we apply the methods to generate the topics and compare the performance of the methods. We then apply a classification method to evaluate the performance of the different topic modeling methods by identifying sections by the topics in the test data. Finally, we use the topic models to classify failed from non-failed bank-year instances in our dataset. We provide a discussion of the results along with several visualizations.

A. Topics Generated by Various Methods

In this set of qualitative experiments, we compare the topics generated by the various methods. Tables II and III show some selected topics generated by NMF, PCA, LDA and KATE for the 8-K and 10-K data, respectively. As for NMF and LDA, we pick the 8 words with the highest probability under that topic. As for PCA and KATE, each topic is represented by the 8 words with the strongest connection to that topic. All of the methods extract meaningful topics from the 8-K and 10-K reports. In the case of 8-K, the three topics shown appear to cover 'shareholder issues', 'management' and 'communications' while for 10-K filings, they cover 'loans', 'financial institutions' and 'cashflow'.

TABLE II
TOPICS LEARNED FROM 8-KS

Topics	Methods	Representative words
share-holder issues	NMF	stockholder, proxy, approval, preferred, record, vote, consideration, matter
	PCA	share, stock, common, dividend, price, issue, security, shareholder
	LDA	vote, meeting, shareholder, proposal, annual, director, approve, election
	KATE	dividend, record, quarterly, declaration, declare, shareholder, stockholder, share
manage-ment	NMF	director, board, committee, member, appoint, require, nonemployee, corporation
	PCA	director, value, security, share, employment, agreement, board, fair
	LDA	director, serve, effective, board, appoint, subsidiary, company, member
	KATE	director, serve, officer, executive, member, age, retirement, position
commu-nications	NMF	presentation, investor, material, make, available, slide, furnish, website
	PCA	shareholder, plan, option, include, security, slide, exhibit, presentation
	LDA	presentation, information, investor, call, conference, exhibit, slide, report
	KATE	presentation, website, conference, slide, available, investor, webcast, management

TABLE III
TOPICS LEARNED FROM 10-KS

Topics	Methods	Representative words
loans	NMF	loan, interest, loss, rate, asset, value, security, income
	PCA	loan, interest, value, rate, asset, loss, income, security
	LDA	value, loan, fair, security, loss, asset, financial, tax
	KATE	loan, stock, reference, incorporate, common, information, dividend, price
financial institutions	NMF	bank, capital, loan, company, institution, hold, financial, deposit
	PCA	bank, institution, capital, company, loan, may, regulation, hold
	LDA	bank, capital, company, institution, financial, hold, banking, deposit
	KATE	bank, institution, capital, hold, regulation, loan, dividend, company
cashflow	NMF	rate, interest, change, income, liability, market, net, risk
	PCA	loss, risk, income, credit, mortgage, share, end, portfolio
	LDA	interest, rate, risk, change, asset, net, liability, income
	KATE	rate, interest, risk, market, simulation, change, net, scenario

B. Visualization of Learned Word Representations

In this section, we visualize the word representations learned by LDA and KATE, in two dimensions, using the TSNE dimensionality-reduction method (t-distributed stochastic neighbor embedding) [37].

For LDA, we simply use the transpose of the topic-word matrix as the word representation matrix. In the case of KATE, each input neuron (i.e., a word in the vocabulary set) is connected to each hidden neuron (i.e., a virtual topic) with different strengths. Thus, each row i of the input to hidden layer weight matrix $\boldsymbol{W} \in \mathbb{R}^{d \times m}$ is taken as an m-dimensional word embedding for word i.

Figures 1 and 2 apply TSNE to visualize the word representations learned by LDA and KATE from 8-K and 10-K reports. We see, from these plots, that KATE-learned word representations possess the highest quality, as many semantically similar or relevant words group together, while dissimilar or irrelevant words are at a distance from each other. We conclude that the capacity of learning meaningful word representations from text is important for evaluating the effectiveness of text analytics techniques.

C. Visualization of Learned Document Representations

A good document representation method is expected to group related documents, and to separate documents from different groups. As stated earlier, we define sections and some subsections of 8-Ks and 10-Ks as documents. We now present visualization for all documents by the topics they represent. Figure 3 shows the TSNE plots of the document representations taken from the 8 main sections of the 8-K reports. We can observe in this comparison that neither NMF nor LDA learn very good document representations, all the points are less distinguishable in the clusters. On the other hand, KATE successfully extracts meaningful document representations for the 8-K reports. It automatically clusters related documents in the same group, and it can easily distinguish between the different sections.

Figure 4 shows TSNE plots for the seventeen major sections and sub-sections of the 10-K reports. Even though the number of sections is significantly higher in 10-Ks, the document representation using LDA and KATE is still good, and indeed much superior to PCA and NMF approaches for topic modeling.

D. Classification Results

We now turn to quantitative experiments to measure the effectiveness of various methods on classification tasks. For classification based evaluation of 8-K and 10-K reports, we train a 2-layer neural network model that uses the learned document representations as input, and directly maps them to the output classes. A softmax classifier with cross-entropy loss was applied for the classification task. Intuitively, high quality representations should produce decent classification accuracies even with a very simple classifier. The classification task is to identify the section or subsection a document corresponds to within the 8-K and 10-K reports, using the topic model representation for the sections/subsections.

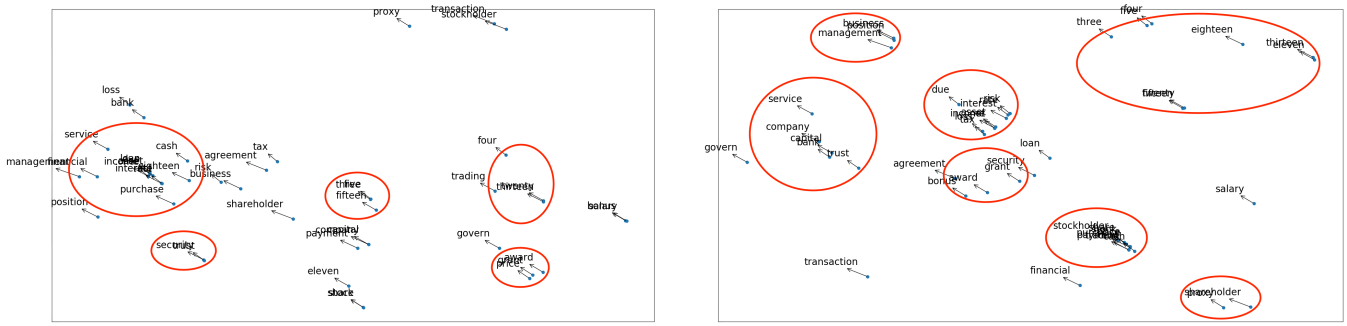
TABLE IV
CLASSIFICATION ACCURACIES ON 8-KS AND 10-KS.

Method	**8-Ks**			**10-Ks**		
	20	**128**	**512**	**20**	**128**	**512**
PCA	0.732	0.809	0.830	**0.843**	**0.912**	0.941
NMF	0.688	0.797	0.816	0.728	0.871	0.899
LDA	0.744	0.808	0.834	0.798	0.874	0.931
KATE	**0.779**	**0.823**	**0.838**	0.821	0.881	**0.947**

Table IV shows the classification accuracy results on the 8-K and 10-K reports in the test set based on different topic modeling approaches. Results are shown for cases when we use different number of topics from 20 to 128 and 512. We can see that KATE is clearly superior in this classification task, and it outperforms the second best model, by a significant margin across various numbers of 8-K and 10-K topics. Surprisingly, even though PCA learns document representations somewhat poorly as shown in Figure 3a, it achieves quite reasonable accuracies. The other observation we make is regarding performance of the topic modeling methods as we increase the number of topics (even larger than the true topic number). Increasing the number of topics/dimensions

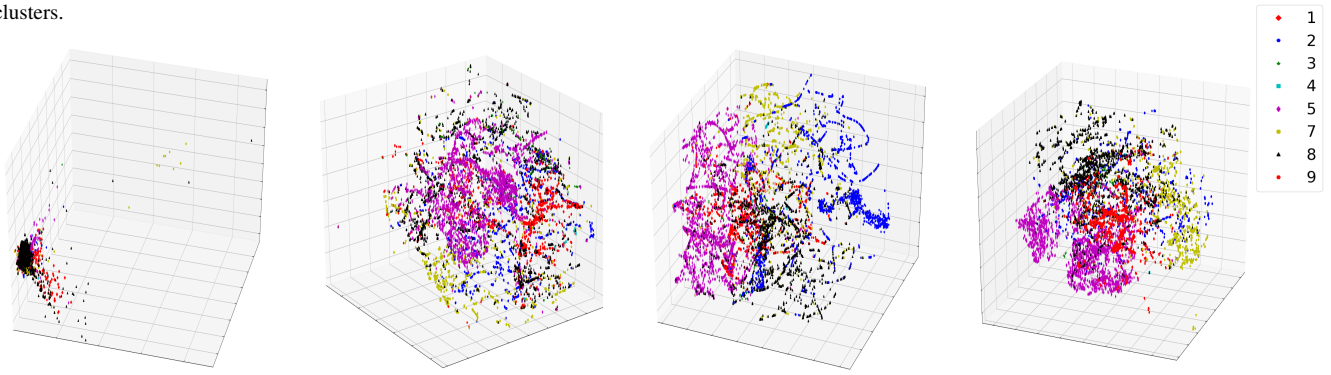
(a) LDA (b) KATE

Fig. 1. 2-D TSNE visualization of 128-D word vectors learned from 8-Ks (using LDA and KATE respectively). The red ovals represent semantically coherent word clusters.



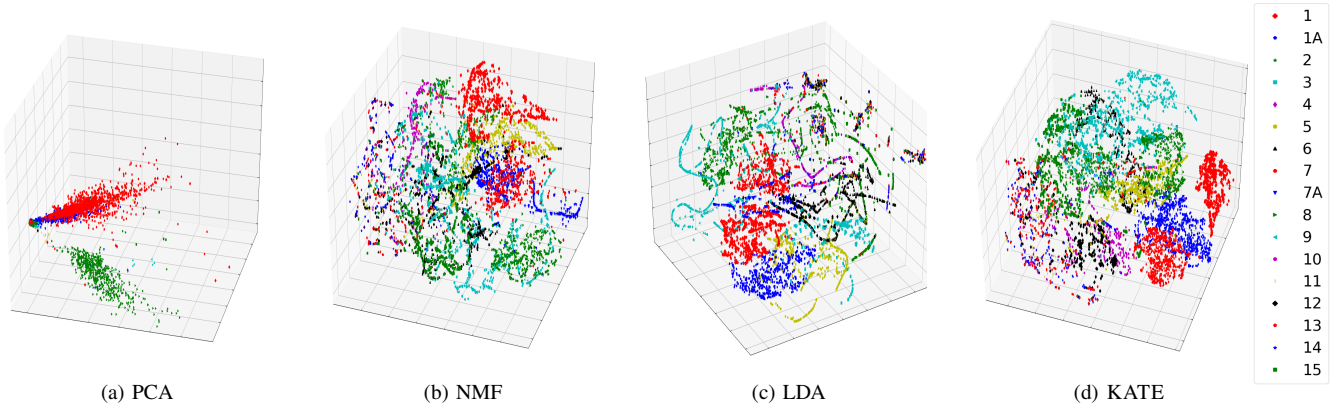
(a) LDA (b) KATE

Fig. 2. 2-D TSNE visualization of 128-D word vectors learned from 10-Ks (using LDA and KATE respectively). The red ovals represent semantically coherent word clusters.



(a) PCA (b) NMF (c) LDA (d) KATE

Fig. 3. 3-D TSNE visualization of 20-D document vectors learned from 8-Ks. (Each document class – section number – is depicted by a different color.)



(a) PCA (b) NMF (c) LDA (d) KATE

Fig. 4. 3-D TSNE visualization of 20-D document vectors learned from 10-Ks. (Each document class – section number – is depicted by a different color.)

improves the classification accuracy. In the case of 10-K data, even though PCA outperforms KATE when the numbers of topics are 20 and 128, KATE achieves a higher accuracy when the number of topics increases to 512.

E. Predicting Bank Failures

Having demonstrated the comparative ability to extract meaningful topics from bank SEC filings using different topic modeling methods, and to learn meaningful representations of these reports, we now extend the analysis to consider the task for categorizing the banks, i.e., whether they failed or not. Different banks, by their various activities, management quality, and risk exposures may have differences in the essential topics and contents of the reports. For example, a bank performing well under a good management and strong investment opportunities would have an emphasis on topics reflecting a positive strategy of growth, while a distressed bank would likely discuss topics regarding challenges in risk control and cost reduction. Taking advantage of the data for failed and non-failed banks from the 2008 global financial crisis, in this section, we analyze how different banks cover different topics in their SEC reports and explore how it relates to bank failures.

We regard the SEC filings of a bank from each year as its bank-year data. We rank the topics for each bank-year data, which indicates the relative weights of different topics covered in the bank-year reports for that bank. Specifically, given the topic vector $\theta_j \in R^K$ of document j learned by one of the topic models utilized in this paper, we compute the topic ranking score vector for document j, where the i^{th} element of the vector is the reciprocal of the ranking (i.e., smaller rank implies higher score) of the i^{th} topic in that document. Here K is the number of topics. We also try introducing a power factor, which actually boosts the performance of LDA and KATE for this bank failure prediction task. Taking an average of the topic ranking score vectors of all the bank-year documents as the topic vector of that bank-year, we use this topic vector as the topical representation of the bank-year and explore how it relates to bank failures.

TABLE V
SIZES OF BANK-YEAR DATASETS

dataset	8-Ks	10-Ks
training set	2,338	1,787
validation set	200	200
test set	1,866	2,164
training set – failed	39	40
validation set – failed	3	4
test set – failed	23	29
training set – non-failed	2,299	1,747
validation set – non-failed	197	196
test set – non-failed	1,843	2,135

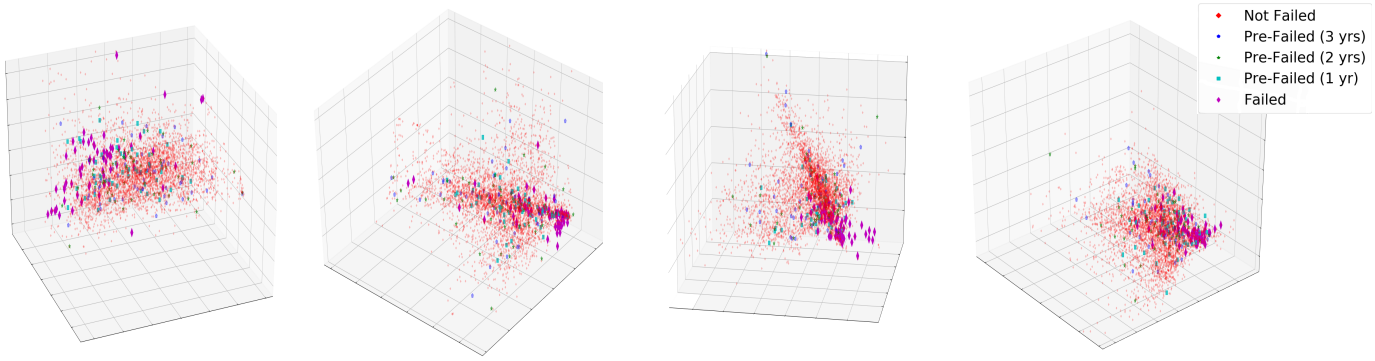
The statistics of the bank-year datasets are summarized in Table V. Note that in this task, in order to make the numbers of failed banks more balanced across the training set and the test set, we use last six years data for testing on 8-Ks and last seven years data for testing on 10-Ks. Note that even though there are 4174 10-K filings (see Table I), we have only 4151 bank-years for 10-Ks, since some banks filed more than once per year. On the other hand, even though there are 50223 8-Ks, there are only 4404 bank years for 8-Ks due to many 8-K filings per bank per year.

Visualization of Bank Vectors: A good bank-year representation using a topic modeling method should group similar bank-years together, and separate different groups. Figures 5 and 6 show the PCA projections of the bank-year representations for the 8-K and 10-K data. It is interesting to note that, for 8-Ks, the failed bank-years share many commonalities in terms of topics as we see that the failed bank-years are all clustered together in the plots. Both the LDA and KATE topic models are effective in capturing this characteristic. Additionally in these plots, we label the bank-years of failed banks one, two and three years prior to the year they failed. For at least a year earlier the topic characteristics of these banks already starts looking dissimilar from the non-failed banks. This provides further support to the hypothesis that a well-constructed topic model can be instructive for bank characteristics assessment. On the other hand, for 10-Ks the topics are not that distinguishable between failed and non-failed banks. We further analyze this issue below; the main problem seems to be the high degree of similarity between entire 10-K filings from both non-failed and failed years for the same bank.

Predicting Bank Failures from a Topic Perspective: Given the results for visualizing bank-year topic characteristics around bank failures, in this section, we again use the cross-entropy loss to train a simple 2-layer neural network model that uses the learned bank representations as input, and directly maps them to the output classes (i.e., bank failure or not) using the softmax function. Table VI shows the classification accuracy results on the 8-K and 10-K bank-year data from different topic models. Interestingly, in this classification task using 8-K reports, LDA is the best method followed by NMF as a close second.

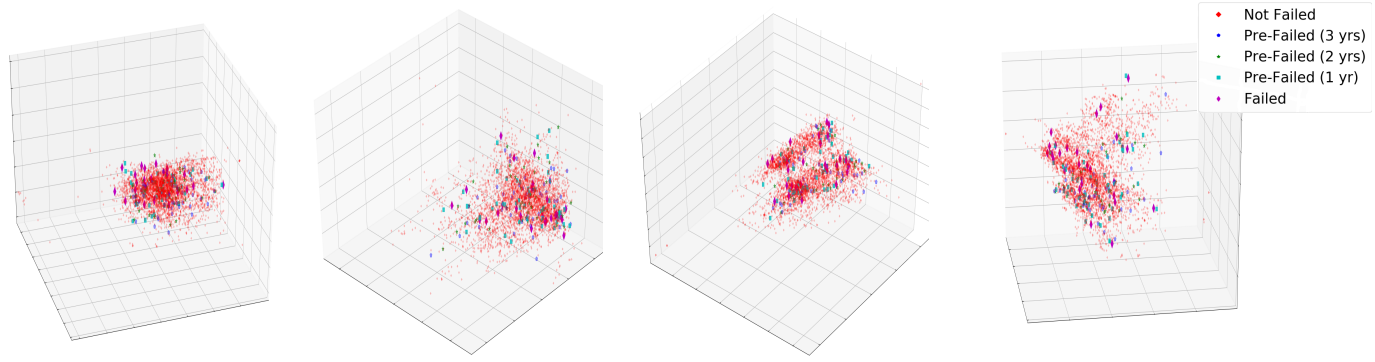
Note that the 10-K reports are extremely imbalanced in terms of bank failure (e.g., there are only 25 failed bank-years among 4,151 bank-years since 27 banks (out of 52) did not file any 10-K on or after their fail years), therefore we extend the definition of the failed year (only on 10-Ks) to also include one year before the exact fail year. In this way, we finally have 73 failed bank-years. Despite the redefinition of a failed bank-year, all our methods failed to work on 10-Ks. We found it is more difficult to predict bank failures using topic models of 10-Ks than those of 8-Ks which can also be verified by comparing the visualization of bank vectors in Figure 5 and 6.

We hypothesize that topic models of 10-K data are less suitable for bank failure prediction because, typically, the sections within a 10-K report inherit most of their content from the previous year. This includes sections 1A and 7 that contain up-to-date discussions relating to the company's financial health and risks. The similarities in textual content between consecutive years can be seen in Table VII. The values in the table are cosine similarities between "bag-of-words" representations of consecutive bank-years. For a bank year, we collect all relevant sections (all 8-K sections, all 10-K sections, all section 1As, or all section 7s) filed by the given bank during the given year. We sum the word counts and create a new bag-of-words representation for the bank year. For each bank holding company, we calculate the median cosine similarity between consecutive non-failed bank-years, and also the similarity between the failed bank-year and the preceding bank-year. The table presents the medians of these per-bank similarity measures. We can observe that the main reason it is hard to discriminate failed versus non-failed years



(a) PCA (b) NMF (c) LDA (d) KATE

Fig. 5. 3-D PCA visualization of 128-D bank-year vectors learned from 8-Ks. (Red diamonds indicate non-failed, while purple diamonds indicate failed bank-years. Blue, green, and cyan points indicate pre-failed bank-years 3, 2, and 1 year(s) before failure, respectively.)



(a) PCA (b) NMF (c) LDA (d) KATE

Fig. 6. 3-D PCA visualization of 128-D bank-year vectors learned from 10-Ks. (Red diamonds indicate non-failed, while purple diamonds indicate failed bank-years. Blue, green, and cyan points indicate pre-failed bank-years 3, 2, and 1 year(s) before failure, respectively.)

using topic-modeling on 10-Ks is the high degree of textual similarity between filings from one year to the next, which is over 0.975, regardless of failed versus non-failed tags. Since all topic-modeling methods are unsupervised, they are unable to extract discriminative topics from the 10-Ks. The median similarity between non-failed and failed years for 8-Ks is only 0.501, which provides enough signal for discrimination between failed and non-failed bank years.

V. Conclusion

In this paper we extend the text mining and analysis beyond a bag-of-words word-frequency based approach. A word frequency based textual analysis runs the risk of getting too conditioned by the relevance and currentness of the word dictionary chosen. Our objective in this paper was to compare and evaluate topic modeling approaches and their effectiveness in analyzing a large set of SEC filings made by US public banks. We applied four major topic modeling methods to a corpus of 8-K and 10-K filings of 578 bank holding companies. The usefulness and effectiveness of topic modeling by the methods considered is evaluated by a few different tasks. The first task was comparing the ability of different topic models to identify the document sections the text was extracted from. Once we obtained favorable results for topic representation of the SEC filings, we examined the ability of the topic models to detect a failed bank from a non-failed bank.

Concurrent to these classification tasks for comparative analytics of the topic modeling methods, we also utilized different visualization methods to examine the topic properties for the banks. We found that KATE method performs the best in terms of distinctness and visualization of sections of the filing reports in terms of topic representation. For classification of sections using the topic models also KATE emerged as the most promising topic modeling method. The final task of classifying banks by topic representation of bank-year showed that while KATE was competitive, the best method turned out to be LDA. Our study shows that novel approaches to topic modeling of financial reports and regulatory filings can be quite instructive for learning bank characteristics, and point the way towards more advance applications of text analytics in financial decision making and risk management.

Acknowledgments

This work was supported in part by NSF grants IIS-1738895 and IIS-1302231.

References

[1] S. J. Alfano, S. Feuerriegel, and D. Neumann, "Do pessimists move asset prices? Evidence from applying prospect theory to news sentiment," in *ECIS*, May 2015.

[2] W. Antweiler and M. Z. Frank, "Is all that talk just noise? The information content of internet stock message boards," *The Journal of Finance*, vol. 59, no. 3, pp. 1259–1294, 2004.

[3] B. Wuthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, and J. Zhang, "Daily stock market forecast from textual web data," in *IEEE Int'l Conf. Systems, Man, and Cybernetics*, 1998.

[4] E. Serrano and C. A. Iglesias, "Validating viral marketing strategies in twitter via agent-based social simulation," *Expert Systems with Applications*, vol. 50, pp. 140–150, 2016.

TABLE VI
FAILED VS NON-FAILED BANK-YEAR CLASSIFICATION

Method	Class	8-Ks (128 topics)				10-Ks (128 topics)			
		precision	recall	f1-score	support	precision	recall	f1-score	support
PCA	F	0.76	0.57	0.65	23	0.00	0.00	0.00	29
	NF	0.99	1.00	1.00	1843	0.99	1.00	0.99	2135
	avg / total	0.99	0.99	0.99	1866	0.97	0.99	0.98	2164
NMF	F	0.94	0.70	0.80	23	0.14	0.03	0.06	29
	NF	1.00	1.00	1.00	1843	0.99	1.00	0.99	2135
	avg / total	1.00	1.00	1.00	1866	0.98	0.98	0.98	2164
LDA	F	0.79	0.83	0.81	23	0.00	0.00	0.00	29
	NF	1.00	1.00	1.00	1843	0.99	1.00	0.99	2135
	avg / total	1.00	1.00	1.00	1866	0.97	0.99	0.98	2164
KATE	F	0.81	0.74	0.77	23	0.00	0.00	0.00	29
	NF	1.00	1.00	1.00	1843	0.99	1.00	0.99	2135
	avg / total	0.99	0.99	0.99	1866	0.97	0.99	0.98	2164

NF: non-failed bank-year. F: failed bank-year.

TABLE VII
THE DEGREE OF TEXTUAL SIMILARITY BETWEEN CONSECUTIVE BANK-YEARS.

Data source	8-K	10-K	10-K §1A	10-K §7
Median of per-bank NF to NF similarity	0.675	0.982	0.965	0.982
Median of per-bank NF to F similarity	0.501	0.975	0.930	0.960

NF: non-failed bank-year. F: failed bank-year.

[5] T. H. Nguyen, K. Shirai, and J. Velcin, "Sentiment analysis on social media for stock movement prediction," *Expert Systems with Applications*, vol. 42, no. 24, pp. 9603–9611, 2015.

[6] G. Ranco, I. Bordino, G. Bormetti, G. Caldarelli, F. Lillo, and M. Treccani, "Coupling news sentiment with web browsing data improves prediction of intra-day price dynamics," *PloS one*, vol. 11, no. 1, p. e0146576, 2016.

[7] F.-T. Tsai, H.-M. Lu, and M.-W. Hung, "The effects of news sentiment and coverage on credit rating analysis." in *PACIS*, 2010, p. 199.

[8] A. Mengelkamp, S. Hobert, and M. Schumann, "Corporate credit risk analysis utilizing textual user generated content-a twitter based feasibility study," in *PACIS*, 2015.

[9] S. Rönnqvist and P. Sarlin, "Detect & describe: Deep learning of bank stress in the news," in *IEEE Symposium Series on Computational Intelligence*, 2015, pp. 890–897.

[10] F. J. L. Iturriaga and I. P. Sanz, "Bankruptcy visualization and prediction using neural networks: A study of us commercial banks," *Expert Systems with Applications*, vol. 42, no. 6, pp. 2857–2869, 2015.

[11] A. Lischinsky, "In times of crisis: a corpus approach to the construction of the global financial crisis in annual reports," *Critical Discourse Studies*, vol. 8, no. 3, pp. 153–168, 2011.

[12] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Systems with Applications*, vol. 41, no. 16, pp. 7653–7670, 2014.

[13] R. P. Schumaker, Y. Zhang, C.-N. Huang, and H. Chen, "Evaluating sentiment in financial news articles," *Decision Support Systems*, vol. 53, no. 3, pp. 458–464, 2012.

[14] T.-T. Vu, S. Chang, Q. T. Ha, and N. Collier, "An experiment in integrating sentiment features for tech stock prediction in twitter," in *COLING*, 2012.

[15] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, "Text mining of news-headlines for forex market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment," *Expert Systems with Applications*, vol. 42, no. 1, pp. 306–324, 2015.

[16] 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.

[17] A. Gupta, M. Simaan, and M. J. Zaki, "When positive sentiment is not so positive: Textual analytics and bank failures," *Available at SSRN 2773939*, 2016.

[18] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *In NIPS*. MIT Press, 2000, pp. 556–562.

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

[20] I. Jolliffe, *Principal Component Analysis*. Springer Verlag, 1986.

[21] Y. Chen and M. J. Zaki, "Kate: K-competitive autoencoder for text," in *Proceedings of the ACM SIGKDD International Conference on Data Mining and Knowledge Discovery*, Aug 2017.

[22] T. Hofmann, "Probabilistic latent semantic analysis," in *UAI Conf.*, 1999, pp. 289–296.

[23] H. M. Wallach, "Topic modeling: beyond bag-of-words," in *23rd international conference on Machine learning*, 2006, pp. 977–984.

[24] X. Wang, A. McCallum, and X. Wei, "Topical n-grams: Phrase and topic discovery, with an application to information retrieval," in *IEEE Int'l Conf. on Data Mining*, 2007.

[25] R. V. Lindsey, W. P. Headden III, and M. J. Stipicevic, "A phrase-discovering topic model using hierarchical pitman-yor processes," in *EMNLP Conf.*, 2012.

[26] C. Wang and D. M. Blei, "Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process," in *Advances in neural information processing systems*, 2009, pp. 1982–1989.

[27] T. Lin, W. Tian, Q. Mei, and H. Cheng, "The dual-sparse topic model: mining focused topics and focused terms in short text," in *WWW Conf.*, 2014.

[28] 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.

[29] X. Wang and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends," in *SIGKDD Conf.*, 2006.

[30] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[31] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[32] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.

[33] A. Makhzani and B. Frey, "k-sparse autoencoders," in *ICLR*, 2014.

[34] SEC, "SEC filings and forms," https://www.sec.gov/edgar.shtml, accessed: 2017-04-17.

[35] I. S. Dhillon and S. Sra, "Generalized nonnegative matrix approximations with bregman divergences," in *NIPS*, vol. 18, 2005.

[36] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *LREC 2010 Workshop on New Challenges for NLP Frameworks*, May 2010, pp. 45–50.

[37] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.