Analyzing Textual Information at Scale*

Lin William Cong Tengyuan Liang Xiao Zhang

First Draft: March 2019; Current Draft: August 2019

Prepared for a book chapter. Comments welcome.

Abstract

We overview recent advances in textual analysis for social sciences. Count-based economic model, structured statistical tool, and plain-vanilla machine learning apparatus each has merits and limitations. To take a data-driven approach to capture complex linguistic structures while ensuring computational scalability and economic interpretability, a general framework for analyzing large-scale text-based data is needed. We discuss recent attempts combining the strengths of neural network language models such as word embedding and generative statistical modeling such as topic modeling. We also describe typical sources of texts, the applications of these methodologies to issues in finance and economics, and promising future directions.

JEL Classification: C55, C80, G10

Keywords: Bag-of-words, Big Data, Machine Learning, Text-based Analysis, Topic Models, Unstructured Data, Word Embedding.

*We are deeply indebted to Jerry Hoberg for his insightful comments. Cong is at Cornell University Johnson Graduate School of Management; Liang is at the University of Chicago Booth School of Business; Zhang is at Analysis Group. Contact author: Cong at will.cong@cornell.edu.

1 Introduction

With the increased capacity of modern computers, it has become feasible to collect enormous amounts of data and then process them through proper aggregation using algorithms to facilitate effective decision-making. For example, financial analysts and investors who used to focus heavily on firms' quarterly earning numbers or infrequent macroeconomic forecasts can now analyze market sentiment using news media articles and forecast business activities with satellite pictures of parking lots.

Big data generally include data of large volume or frequency, data from non-conventional courses, and unstructured data that require special processing and information extraction. Texts are a pre-dominant form of unstructured data and there is as much information in language data as there is in numbers, not to mention the greater interpretability texts offer. They enable econometricians to supplement or replace traditional surveys, capture more granular and up-to-date information, and complement information extracted from structured data such as financial ratios. However, it has been challenging to analyze texts at a large scale and in a way that preserves interpretability.

We therefore aim to help future researchers to understand the important recent developments and applications in the field of textual analysis and see how computing capacity can help them utilize textual data. It is absolutely crucial to build algorithms to aggregate data and extract information to facilitate any decision we may need to make. In this article, we discuss several approaches to this end, and highlight their strengths and weaknesses.

Analyzing textual data is challenging for several reasons: first, language structures are often too intricate and complex to be summarized by simply counting words; second, textual data are high-dimensional and processing a large corpus of documents is computationally demanding; third, there lacks a framework relating textual data to sparse regression analysis traditionally used in social sciences while maintaining interpretability.

Applying textual analysis in financial markets and business environments is even more challenging than in other fields because they evolve faster than physical laws or genetic codes,

which means predictive models built on past data are not sufficient without economic understanding and interpretability.[1] In fact, one should recognize that scientists and practitioners use textual analysis primarily because texts offer more interpretability.

As such, we focus on information richness, computational efficiency, as well as economic interpretability when assessing various methodologies for textual analysis. In what follows, we first discuss typical sources of textual data and then the current approaches to textual analysis in social sciences, statistics, and machine learning fields respectively. We do not claim to do full justice to the literature because this is not a survey of all relevant studies, but instead aim to illustrate major themes in recent developments.[2]

2 Texts as Unstructured Data

Textual data manifest themselves in various forms. Here we list textual data that are easily available to researchers and decision-makers. The list is necessarily partial, with an emphasis on data related to economics and finance. Our goal is to illustrate what kind of data sources prove to be useful for textual analysis.

News. The Wall Street Journal's (WSJ) data are widely used in various academic studies and particularly suitable to textual analysis. We focus on front-page articles only because these are manually edited and corrected. This is particularly useful for newspaper in earlier years as they are scanned and digitized using OCR (optical character recognition), which tends to generate typos.

Other newspapers, such as the New York Times, the Financial Times, and the Economist, contain relevant information in the Economic, Business and Finance sections. They are

[1] The signal-to-noise ratios in economics or finance settings can also be much lower than those in scientific or engineering settings. The data generation process is also typically non-experimental.

[2] There are excellent surveys on textual analysis, including Li (2010) on manual-based textual analysis and past topics and future directions, Kearney and Liu (2014) on textual sentiment, Das et al. (2014) on basic code snippets and basic text analytics. In particular, Loughran and McDonald (2016) underscore how textual analysis is substantially less precise and that understanding the art is of equal importance to understanding the science.

available from, for example, Proquest (`https://www.proquest.com`).

Firm-specific news from Factiva (`https://www.dowjones.com/products/factiva`) is also a great resource if cross sectional variation is more important for a particular research question. This firm-specific news enables us to explore variation in texts among firms in the cross-section.

Corporate filings and releases. Company filings are typically available for public firms. For example, they have been publicly available in the United States since 1993. To facilitate the rapid dissemination of financial and business information about companies, the U.S. Securities and Exchange Commission (SEC) allows publicly-listed firms to file their securities documents with the SEC via the Electronic Data Gathering, Analysis and Retrieval (EDGAR) system (`https://www.sec.gov/edgar/`). We discuss in this article several frquently used forms, such as the Management Discussion and Analysis (MD&A) sections of the annual report (10-K), IPO prospectus (S-3), and current reports (8-K).

1. Management Discussion and Analysis (MD&A). MD&A is a section of a public company's annual report (10-K) or quarterly filing (10-Q), in which the management analyzes the company's performance with qualitative measures. Since this section is unaudited, management has the most discretion and flexibility in terms of creating its content. Typically, MD&A provides commentary on financial statements, systems and controls, compliance with laws and regulations, financial activities, and actions it has planned or has taken to address any challenges the company is facing. Management also discusses the firm's outlook by analyzing industry trends, competitive environment, economic conditions, and risks in the financial market.

2. Risk Factor Discussions. In Section 1A of 10-K reports, companies discuss potential risk factors associated with business and financial operations. According to Regulation SK (Item 305(c), SEC 2005), firms are legally obliged to disclose "the most significant factors that make the company speculative or risky." Therefore, typical risk factors discussions include local economic, financial, and political conditions, government regulation, business li-

censing or certification requirements, limitations on the repatriation and investment of funds and foreign currency exchange restrictions, varying payable and longer receivable cycles and the resulting negative impact on cash flows. Since companies may get sued if they do not warn investors and potential investors about potential risk, firms tend to include many risk discussions that are only remotely relevant to them.

3. Proxy Statements. Firms need to file a proxy statement (DEF 14A) in advance of the annual meeting to provide shareholders with sufficient information about upcoming meetings, whenever they hold shareholder meetings and solicit votes. A proxy statement often includes information on shareholder proposals, voting procedures, background information (including potential conflicts of interests) of nominated directors, compensation structure of board and executives, and auditors. Most shareholder proposals that are up for votes are approval of the re-election of directors, approval of executive compensation plan, approval of audit fees, and ratification of the ongoing engagement of the auditing firm.

4. Conference Call or Meeting Transcripts. Most publicly traded firms hold regular conference calls with their analysts and other interested parties. During the conference call, management gives its view on the firm's past and future performance and responds to questions from call participants. Both audio recordings and transcripts for conference calls are available. For example, one can obtain conference-call transcripts from SeekingAlpha (`https://seekingalpha.com/`).

Another meeting transcript often used is from the Federal Open Market Committee (FOMC) meetings. (`https://www.federalreserve.gov/monetarypolicy/fomc_historical.htm`). Every year, the FOMC holds eight regularly scheduled meetings. FOMC meeting members discuss the economic outlook and formulate monetary policy during these meetings. All policy changes are made public in a short meeting statement immediately after the meeting. In addition, detailed records of the discussions during each meeting (minutes) are released a day later.

5. Analyst Reports. Analyst reports from Investext via Thomson One (`https://www.`

thomsonone.com/). Equity analysts from major investment banks periodically write about firms' past performance and their view about firms' future stock price.

6. Patents. Recently, the United States Patent and Trademark Office (USPTO) has made their patent data public available (https://bulkdata.uspto.gov/). This includes textual data, such as patent application and grant. Each patent documents consists of both abstract and detailed description, as well as citations. The data goes back as early as the 1920s and covers essentially all patents filled with USPTO. This greatly reduces the workload for researchers who want to use this type of patent data. Previously, researchers needed to scrape patent documents from Google Patent, which could be more labor intensive and time consuming. On the other hand, what's unique about Google Patent is that it collects patents from 100+ patent offices around the world, making the coverage much broader.

3 Count-based Analyses in Economics and Finance

Count based methods are generally easy to interpret because researchers who have domain knowledge define the bag of words or the dictionary. This line of studies count the occurrence of pre-specified keywords and phrases as a way to summarize information in the text.[3] One example is Nini, Smith, and Sufi (2012) that examines the impact of covenant violations on corporate behavior. Given that there is no existing database on covenant violations, the authors identify covenant violations by applying a simple textual analysis methodology to firm filings. What they do is to search for the keyword "covenant" in 10-k filings (annual reports). Conditional on finding this keyword, their algorithm then searches for additional keywords, such as "waiv," "viol," "in default," "modif," and "not in compliance," within three lines above or below the line containing "covenant" to make sure the texts indeed discuss covenant violations.

This is a typical application of count-based textual analysis to an economic question. The research question is particularly important in the economics and finance literature;

[3]Antweiler and Frank (2004) pioneered attempts to utilize textual information in economics and finance.

while there is plenty of theoretical work studying the role of creditors on the governance of corporations both inside and outside of bankruptcy, there are very few empirical studies done on a large scale due to the limited availability of structured covenant violation data. After building a data set on covenant violations using texts, the authors use it to study the effect of credit control rights. They find that covenant violations are prevalent and are followed immediately by declines in acquisitions and capital expenditures, sharp reductions in leverage and shareholder payouts, and increases in CEO turnover.

The authors are cognizant of the shortcomings of the count-based method: it requires domain knowledge and significant manual work. Indeed, the authors go through a large number of iterations to pin down a list of best keywords. Despite such efforts, the method produces a large number of false positives. To make the data as clean as possible, the authors also hired a group of research assistants to manually go over the filings to eliminate the false positive. This step is almost inevitable for most research in finance and economics employing count-based methods.

Another salient example is Baker, Bloom, and Davis (2016), which constructs a new index of economic policy uncertainty (EPU) based on newspaper coverage frequency. Their approach is straightforward: search and count the occurrence of some keywords related to economic uncertainty in 10 leading US newspapers. The keywords include "economic" or "economy"; "uncertain" or "uncertainty," "congress," "deficit," "Federal Reserve," "legislation," "regulation," and "White House," etc. What they highlight in their paper is an extensive audit study of 12,000 randomly selected articles drawn from major US newspapers. The auditors manually assess whether a given article discusses economic policy uncertainty, which not only lends credibility to their results but also provides insights on domain knowledge for future studies.

Yet one more important application of textual analysis in social science entails extracting sentiment information from texts. In a pioneering study, Tetlock (2007) uses General Inquirer's Harvard IV-4 psycho-social dictionary as a keyword list and counts the number of

words in each day's WSJ that fall within various word categories. Employing principal components factor analysis, he extracts the most important semantic component to construct media sentiment. He finds that high media pessimism predicts downward pressure on market prices followed by a reversion to fundamentals, and unusually high or low pessimism predicts high market trading volume. A number of follow-up studies document similar results as Tetlock (2007).

These types of asset pricing prediction exercises are often based on one of the two assumptions. First, the stock market may be inefficient, meaning that not all information in the newspapers is incorporated in the stock market. Second, newspapers not only report on the state of the economy but also play an active role in influencing it. While the first assumption is difficult to justify, some studies successfully validate the second assumption. For example, Wisniewski and Lambe (2013) use pre-defined word lists to measure the intensity of negative media speculation and show that negative media attention of the banking sector has real effects. They show that over the sub-prime crisis pessimistic coverage Granger-caused the returns on banking indices, while causality in the opposite direction is not as significant, which suggests journalistic views have the potential to influence market outcomes. The caveat is that their sample period is an extreme state of the world, i.e., the Great Recession, and the results may not apply to other less extreme states of the economy. This is partially why most results in this strand of literature are driven by observations in the recessions.

For the count-based approaches, domain knowledge is especially important. In other words, coming up with keywords that suit the specific application is crucial. There are more and more dictionaries available to researchers, and some of these dictionaries are constructed to suit research specifically in economics and finance. Loughran and McDonald (2011) document that some widely used dictionaries do not work well in the finance context. Their results highlight the importance of domain knowledge. To solve this problem, they constructed a refined wordlist that applies to financial contexts, and also made it publicly available for others to use. More specifically, they classify words into negative, positive, uncertainty, liti-

gious, strong modal, weak modal, constraining categories. They show that predictive power increases using their keyword list over other more general-purpose dictionaries.

Since then, Loughran-McDonald Sentiment Word List has become one of the most widely used dictionaries in finance research. As of April 2019, there are more than 1,900 citations for Loughran and McDonald (2011) and many of those citations are from academic research that uses their wordlist. Our view is that there will be more and more refined dictionaries coming out, aiding researchers taking simple count-based approach to analyze text data. At the same time, tools developed in other fields could prove to be useful. For example, Bollen, Mao, and Zeng (2011) using other dictionary-based tools such as OpinionFinder and Google's Profile of Mood States to measure sentiment in Twitter messages and correlate it with stock market movements.

Pre-defining dictionaries or manually labeling as done in the aforementioned studies could perform well. However, such an approach has several limitations. First, because the way of defining the dictionary or bag of words is task-specific and requires domain expertise, count-based methods (at least with static dictionaries) thus may not be generalizable or flexible as an analytics tool for studying a wide range of problems. It achieves scalability once variables are guided or constructed by the researchers. But to select the right model and construct the variables, researchers also have searched over a complex space which is computationally expensive, and domain knowledge takes years to accumulate. Second, with a large dictionary, representing each word as a long vector with all but one entry being non-zero means the computations are inherently high-dimensional. Third, count-based methods leave out finer linguistic structures and may miss important information in the texts.

For additional survey articles on text-based analysis in economics, sociology, and political science, please see Gentzkow, Kelly, and Taddy (2017), Evans and Aceves (2016), and Grimmer and Stewart (2013) survey. In particular, Gentzkow, Kelly, and Taddy (2017) point out that new techniques are needed to deal with the large-scale and complex nature of textual data.

4 Statistical Inference and Regression Models

Beside count-based methods that require domain knowledge, data-driven and model-based inference has become increasingly popular for analyzing textual information for decision-making. Latent Dirichlet Allocation (LDA) as one of the topic modeling techniques is arguably the most widely used one. LDA discovers the abstract topics that occur in a collection of documents, and in so doing classifies texts document to different topics.

LDA models assume a simple, two-distribution data generating process where each document is generated from a (latent) distribution over a collection of topics and each topic is a distribution over the words in the vocabulary. LDA proposes a hierarchical Bayesian model for the generative process of each document d. First, each topic $\beta_k \sim \text{Dirichlet}(\eta)$ is a multinomial distribution over the vocabulary of words. Second, one generates a multinomial distribution over K-topics for this particular document d, denoted as $\theta_d \sim \text{Dirichlet}(\alpha)$. The word generation process for this document d is as follows: for word W_{di} in this document, sample a specific topic $z_{di} \in \{1, 2, \ldots, K\}$ with $z_{di} \sim \theta_d$, then sample the observed word $W_{di} \sim \beta_{z_{di}}$. Or equivalently, the probability of word W_{di} to be word w in the dictionary is

$$P(W_{di} = w|\theta_d, \beta_1, \ldots, \beta_K) = \sum_k \theta_{dk}\beta_{kw} =: [\Theta B]_{dw} \ ,$$

where the matrix notation $\Theta := [\theta_1, \ldots, \theta_D]' \in \mathbb{R}^{D\times K}$ and $B = [\beta_1, \ldots, \beta_K]' \in \mathbb{R}^{K\times V}$.

Here we provide a simple illustration. Suppose we have the following set of text documents. Each text contains only one sentence.

Text 1: Economics studies the behavior and interactions of economic agents.

Text 2: Microeconomics, Macroeconomics, and econometrics are the most prominent fields in economics.

Text 3: Education is the process of acquiring knowledge, skills, and values.

Text 4: Formal education includes many stages, such as preschool or kindergarten, elementary school, high school, college, and graduate schools.

Text 5: Economic training is an essential part of the curriculum in many stages of education; most high schools offer courses on microeconomics and macroeconomics.

LDA could produce something such as the following.

Text 1 and 2: 100% Topic A

Text 3 and 4: 100% Topic B

Text 5: 60% Topic A, 40% Topic B

Topic A: 70% economics, 10% microeconomics, 10% macroeconomics, 10% econometrics.

Topic B: 40% education, 30% school, 10% knowledge, 10% skills, 10% values.

It is then up to the researchers to interpret the topics. In this example, Topic A could be interpreted to be about economies whereas B to be about education.

At a high level, the algorithm of LDA is as follows. First, researchers specify the number of topics, K, in the collection of the documents. Second, for words in the corpus, LDA randomly classifies each word as one of the K topics. Third, suppose that the topic assigned to a word is wrong but the topics assigned to other words are correct, then assign another topic to this particular word. We choose the new topic based on the topics in this document and times this word assigned to other topics in all of the documents. Finally, repeat this process a number of times for each document and calculate the relative weight of each topic.

Since the LDA model has been around for the past decade, there are many LDA packages written in many statistical languages and are very easy to use. One just needs to clean and tokenize the text data before feeding them into LDA packages. Researchers have applied LDA to analyzing all sorts of text data in finance. For example, Huang, Lehavy, Zang, and Zheng (2017) apply LDA to compare conference call transcripts and subsequent analyst reports; Jegadeesh and Wu (2017) study the information content of Federal Reserve communications; Hansen, McMahon, and Prat (2017) also analyze FOMC meeting transcripts during Alan Greenspan's tenure and find transparency leads to greater accountability; Hassan, Hollander, van Lent, and Tahoun (2017) use LDA on firms' quarterly earnings conference calls transcript to construct a new measure of political risk faced by individual US firms; Bandiera, Lemos,

Prat, and Sadun (2017) examine a large panel of CEO diary data and uncover two distinct behavioral types.

It has been observed that LDA becomes very computationally expensive on large data sets (Mikolov, Chen, Corrado, and Dean, 2013), and without principled prior choices extremely common words tend to dominate all topics (Wallach, Mimno, and McCallum, 2009). Therefore, applying LDA directly to financial or economic settings with big data could be ineffective or even misleading.

While topic modeling is widely used in the finance literature these days, it is not the only methodology in this category. Manela and Moreira (2017) take a regression approach to construct an index of news-implied market volatility based on text from the WSJ from 1890-2009. They apply support vector regression, which uses a penalized least squares objective to identify a small subset of words whose frequencies are most useful for matching patterns of turbulence in financial markets. They find that news coverage related to wars and government policy explains most of the time variation in risk premia their measure identifies. Kelly, Manela, and Moreira (2018) develop an economically motivated high dimensional selection model that improves machine learning from texts. The model adapts the selection model of Heckman (1979) and consists of two components that are tailored to the twin challenges of high dimensionality and sparsity of text data. The first component is the selection equation. This models the text producer's choice of whether or not to use a particular phrase. The second component is a positive counts model, which describes the choice of how many times a word is used (conditional on being used at all). They use their framework to perform various backcast, nowcast, and forecast exercises. Gentzkow, Shapiro, Taddy, et al. (2016) measure trends in the partisanship of congressional speech from 1873 to 2016, defining partisanship to be the ease with which an observer could infer a congress-person's party from a single utterance. The authors adopt two estimation approaches. The first is a leave-out estimator that addresses the main source of finite-sample bias while allowing for simple inspection of the data. The second, our preferred estimator, uses a lasso-type penalty on key model

parameters to control bias, and a Poisson approximation to the multinomial logit likelihood to permit distributed computing.

Several other models specifically deal with the ultra-high dimensional nature of the text documents. For example, Taddy (2015) approximates the multinomial distribution of each word with independent Poisson regressions. His model is clever in the sense that the Poisson regression can be distributed across parallel computing units, making the implementation computationally feasible. Kelly, Manela, and Moreira (2018) further extend this model to make it more applicable to economics and finance context. They build on Taddy's distributed Multi-dimensional Regression (DMR) insight of independent phrase-level models but replace each phrase-level Poisson regression with a hurdle model that has two components. The first component is a selection equation, which models the text producer's choice of whether or not to use a particular phrase (similar to the idea in Heckman, 1979). The second component is a positive counts model, which describes the choice of how many times a word is used. The idea behind their model is that not only do positive phrases count as informative, but whether some words are used at all also conveys important information.

5 Machine Learning and NLP

Recent tools from the natural language processing (NLP) literature presents an alternative for analyzing textual data. Machine learning techniques such as neural networks language models preserve the syntactic and semantic structure well while maintaining computational tractability. Yet these models are often not transparent and thus are limited in their direct applications in social sciences, which often require economic inference and interpretation. In fact, they are often referred to as "black box" models in statistics.

Word embedding is arguably the most popular representation of document vocabulary within the NLP category. It captures context of a word in a document, semantic and syntactic similarity, relation with other words, etc., via representing words in vectors. Compared to the count-based method, word embedding models are data-driven. The idea is that words

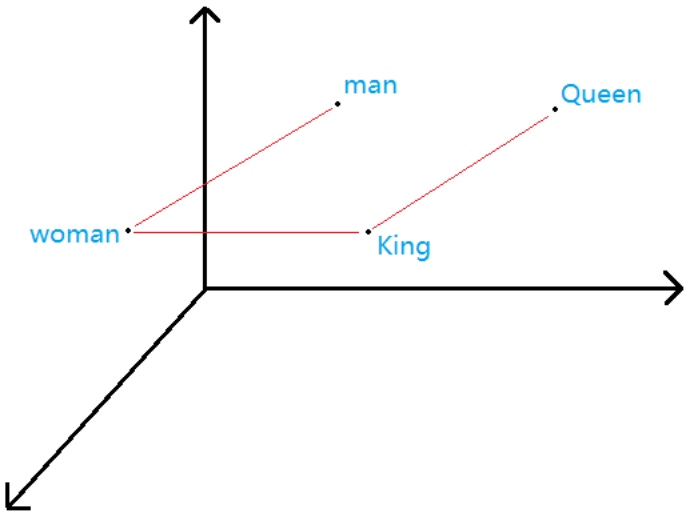

Figure 1: A Graphical illustration of King/Man - Woman/Queen Example

tend to co-occur with neighboring words with similar meanings. In the vector space, words are relationally oriented in the sense that words with similar meaning are closer to each other. In addition, distances between words turn out to have meanings as well. The most famous example is "King/Man - Woman/Queen" relationship. Taking vector("King") vector("Man") + vector("Woman") results in a vector that is closest to the vector representation of the word Queen.

Neural networks are actually not new, they have been around for decades but the lack of accessible, affordable computational power as well as available data was a major bottleneck. The advent of more sophisticated algorithms, computational powers from GPUs becoming cheaper and data literally flooding in from all sources have led to what can be called a renaissance for deep learning. The major advantage of these models over traditional models is the performance gain with the increase in the amount of data. Neural-network-based models become better and better as the data size increases.

Many recent studies argue that vector-based representation exhibits both syntactic/semantic and computational advantages over the classic index-representation and count-based methods. Based on developments in the state-of-the-art neural network language models, Mikolov, Chen, Corrado, and Dean (2013) (word2vec) proposed simple network architectures that

learn high-quality high dimensional vector representation of words from huge datasets. One key advantage of the word vector representation is that it measures multiple degrees of similarities both in the syntactic and semantic sense, and that similar words are "close" to each other in the vector representation. There are two main approaches for learning semantic vector representations.

Bengio, Ducharme, Vincent, and Jauvin (2003); Mikolov, Chen, Corrado, and Dean (2013); Mikolov, Sutskever, Chen, Corrado, and Dean (2013) propose one-hidden-layer neural networks models to learn the representation (word2vec). The hidden layer (with p hidden units) encodes the vector representation $w, \tilde{w} \in \mathbb{R}^{p \times V}$.[4] Then based on *local* context windows, one aims to optimize $w, \tilde{w}$,

$$\min_{w,\tilde{w}} \quad - \sum_{\substack{i \in \text{corpus} \\ j \in \text{context}(i)}} \left\{ \langle w_i, \tilde{w}_j \rangle - \log \left(\sum_{k \in V} \exp(\langle w_i, \tilde{w}_k \rangle) \right) \right\} \quad .$$

Mikolov, Sutskever, Chen, Corrado, and Dean (2013) subsequently propose computationally efficient approximation schemes including "negative sampling" and "hierarchical soft-max" to train word2vec models, which scale well with tasks involving billions of words (Mikolov, Chen, Corrado, and Dean, 2013).

Another representation learning approach is proposed by Pennington, Socher, and Manning (2014) and is based on *global* concurrence X_{ij}. For some pre-chosen weights function $f(\cdot)$, one optimizes the weighted least squares to learn $w, \tilde{w} \in \mathbb{R}^{p \times V}$

$$\min_{w,\tilde{w}} \quad \sum_{i,j \in V} f(X_{ij}) \left(\langle w_i, \tilde{w}_j \rangle - \log X_{ij} \right)^2 \quad .$$

A good choice of $f(x) = (x/x_{\max})^{3/4} \wedge 1$, see Pennington, Socher, and Manning (2014).

Simply put, Word embedding aims to represent words via vectors such that similar words

[4]Here we focus on the skip-gram model to predict its context based on a word, w corresponds to weights between the input layer and the hidden layer, and $\tilde{w}$ denotes the weights between the hidden layer and the output layer.

or words used in a similar context are close to each other while antonyms end up far apart in the vector space. Contrary to count-based methods, these vectors are dense (generally a few hundred dimensions as opposed to the number of unique words in all text documents).

Word2Vec is one of the most popular methods to construct word embedding representation. There are two algorithms that generate word2vec embeddings, CBOW and Skip-Gram. Given a set of text documents, the model loops on the words of each sentence and either tries to use the current word to predict its neighbors (its context), in which case the method is called Skip-Gram, or it uses each of these contexts to predict the current word, in which case the method is called Continuous Bag of Words (CBOW). Both algorithms yield satisfying results.

Most applications of word2vec or other word embedding models are in computer science, such as automatic summarization, machine translation, named entity resolution, sentiment analysis, information retrieval, speech recognition, and question answering. It is still relatively new to researchers in economics and finance, though we do believe that there will be more and more papers that capitalize on the advantages of these models.

Li, Mai, Shen, and Yan (2018) is an elegant example applying word2vec in finance. The authors first learn the meanings of all the words and phrases from earnings call transcripts. They then construct a "culture dictionary" of words and phrases culled from earnings call transcripts that most frequently appear in close association with each five cultural values: innovation, integrity, quality, respect, and teamwork. They find that corporate culture plays significantly influences deal incidence and merger pairing, and that post-merger, acquirers' cultural values are positively related to their target firms' cultural values pre-merger.

One challenge facing these applications of word embedding is interpretation, both interms of model complexity/transparency and economic explanability. The embedding naturally introduces notions of distance among vector representation of words or phrases, one can therefore potentially use clustering techniques to further enhance the interpretability of word groups.

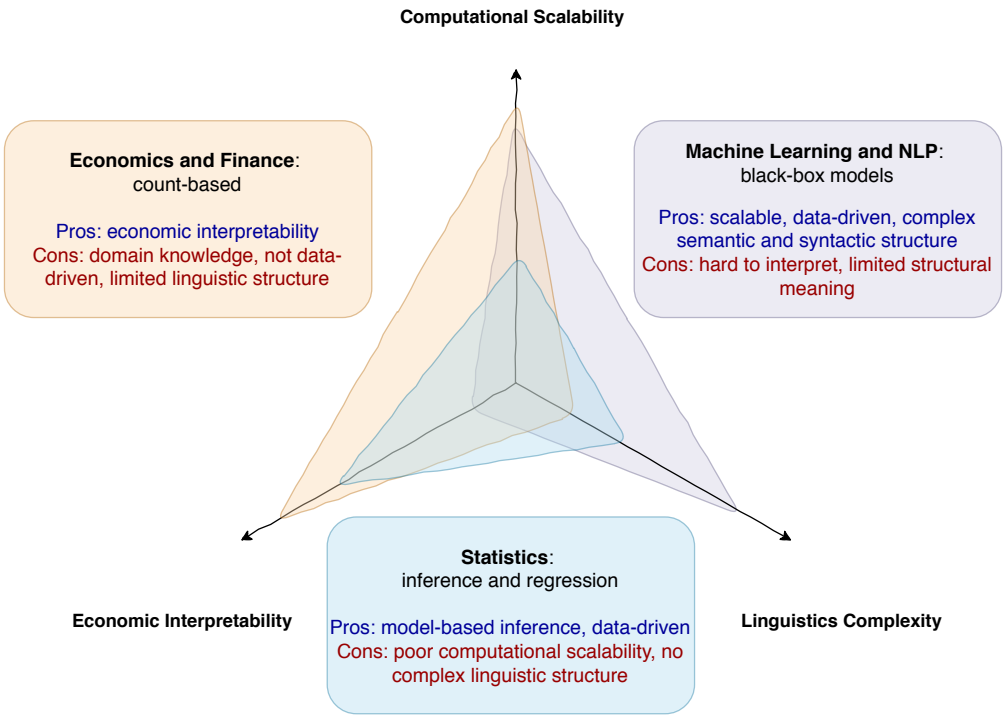

Figure 2: Tradeoffs in Various Approaches. Reproduced from Cong, Liang, and Zhang (2018), Figure 1.

6 A Textual-Factor Framework

The various tradeoffs in the above three approaches are summarized in Figure 2. Given the speed at which rich textual data are generated and the fast pace of developments of industry applications, many attempts have been made recently to analyze texts at a large scale while allowing information richness and ensuring computational efficiency and economic interpretability. We elaborate on one attempt entailing the use of "textual factors."

For example, Cong, Liang, and Zhang (2018) develop a textual-factor framework to potentially tackle problems encountered in current approaches. The authors draw insights and strengths from both neural network models for natural language processing and topic models in statistical machine learning. In particular, they develop a framework to summarize and analyze textual data, with the goal of preserving the informational structure (syntactic and semantic) encoded in natural languages, ensuring computational scalability and economic interpretability, and relating to linear regression models commonly used in social sciences.

They then demonstrate the efficacy of the textual factors generated and apply them to study issues in finance and economics.

Their textual-factor approach involves two stages. First, they form an interpretable set of vectors from the textual data that "spans" the word-document space. In other words, the authors identify a small number of textual factors that explain main variations in the texts. Second, they project each data sample of texts onto the textual factors to find out the beta loadings, which are quantitative measurements/explanatory features for downstream regression tasks.

The goal of the first stage is to generate textual factors to adequately represent the textual data, allow fast computation, and preserve interpretability. It further comprises of three steps, as we describe next.

(1a) Word Embedding They start with a continuous vector embedding of each word in a large vocabulary using neural networks (word2vec) in order to construct the semantic and syntactic links of words in the texts. This step represents words or multi-grams in the texts in a way to capture the rich information and complex language structure.

Count-based and statistical models for textual analysis in social sciences traditionally adopt the "one-hot" representation: words (or N-grams) are treated as very high dimensional vectors/indices over a vocabulary with only one 1 and lots of 0's. Such approaches leave out any consideration of the semantic relations among words, and therefore lack natural notions of similarity among words, resulting in sparse, high-dimensional, and noisy representations.

In contract, Cong, Liang, and Zhang (2019) use semantic vector representations obtained in the NLP literature, which account for word similarities, preserve the language structure and reduce ambient noise. Specifically, each unique word is mapped to a real-valued p-dimensional vector, where $p \ll V$. The dimensionality p of the real-valued vectors can be orders of magnitude smaller than the dimensionality V of the "one-hot" representation.

(1b) Scalable Clustering The authors build on the vector representation to cluster vectors pointing into similar directions. The second step is a key innovation and aims at balancing the interpretability and complexity of their model, and reducing dimensionality for computational ease. As "educated guesses" of the true topics, the clusters would be used for the third step of topic modeling.

The semantic vector representation significantly reduces the dimensionality compared to the classic "one-hot" representation. However, to capture the language structure well, the representation is still inherently high dimensional (with p typically being few hundreds). In addition, the total number of words in the vocabulary is oftentimes very large (V at least ten thousands for real applications). Since the semantic vector representation preserves similarity, the authors argue that a natural next step is to cluster words that are similar to each other through unsupervised learning, yielding in a data-driven way a number of "topics/clusters" that are easy-to-interpret.

That said, clustering in high dimensions is notoriously hard both statistically and computationally. For most classic clustering methods, the computation complexity ($O(V^2p)$ in our case) depends on the number of items to cluster (denoted as V), which has poor scalability in practice. To overcome this challenge, the authors resort to the latest theoretical computer science literature and apply the so-called locality-sensitive-hashing (LSH) (Datar, Immorlica, Indyk, and Mirrokni, 2004; Andoni, Indyk, Laarhoven, Razenshteyn, and Schmidt, 2015) in our setting.

The basic idea behind LSH it to return near-neighbor information in near-linear time through constructing a family of hash functions H with the following property: for a random element $h(\cdot) \in H$

$$
\begin{aligned}
&h(x) = h(y) \text{ with probability at least } 1 - p_1, && \text{for any } x, y \text{ such that } d(x, y) \leq d_1, \\
&h(x) = h(y) \text{ with probability at most } p_2, && \text{for any } x, y \text{ such that } d(x, y) \geq d_2,
\end{aligned}
$$

where the probability is with respect to the sampling of the hash functions.[5] Intuitively, the hash functions help assessing similarity in that they seldom claim two items to be similar when they are actually far away, nor do they conclude two close items to be disparate.

Building upon the LSH technique for approximate near neighbor search, the authors introduce several scalable clustering algorithms. They then demonstrate the quality of the clusters and the scalability of the methodology.

(1c) Guided Topic Modeling The authors use the clustering results obtained in (1b) to guide and enhance a topic model. Because LDA is computationally expensive on large data sets and lacks separability, the authors advocate a "clustering" perspective of topic modeling. Much of the statistical and computational difficulty for topic modeling roots in the fact that LDA allows topics to have overlap in terms of words, rather than separability (or, "anchor words," meaning words that only appear in one unique topic). Without the separability of topics, it is very hard to clearly identify various topics (provably NP-hard, Sontag and Roy, 2011). They overcome this difficulty by learning the separability of topics in a data-driven way by incorporating the semantic vector representation. That is, they utilize the vector representation of words as guidance and enhance our topic modeling approach. Based on the semantic similarity among words captured by the vector representation, it is more likely that close-by words belong to the same topic. This prior knowledge significantly reduces the search space/complexity of the topic-word distributions, therefore easing the optimization approach.

With the word clusters obtained earlier, they then develop computationally efficient and conceptually simple methods to learn textual factors. Because the word lists of topics are more disjoint, the topics are largely distinct from each other, in contrast to the case in plain-vanilla LDA where extremely common words (or, stop words) dominate multiple topics (Wallach, Mimno, and McCallum, 2009). The key computational trick is then to estimate one topic at a time given the separability of clusters. For instance, given the i-th cluster,

[5]Near-linear time means $O(Vk)$ where k is the number of hash functions to generate the Hash table.

with support (set of indices) $S_i \subset [V]$, we can focus on the document-term submatrix N_{S_i} where the columns consist of words only in i-th cluster. In the paper the authors implement this procedure using a frequentist-approach to topic modeling, i.e., Latent Semantic Analysis (LSA) through Singular-Value Decomposition.

The authors claim that such data-driven guidance significantly enhances the performance of the topic model for unsupervised learning. More empirical work can test this claim.

(2) Beta Loadings on Textual Factors Suppose that from the first stage we obtain K textual factors, where K is endogenously specified and can potentially be data-dependent. The set of textual factors are then represented by the triplet $(S_i, F_i \in \mathbb{R}^{|S_i|}, d_i \in \mathbb{R}_{\geq 0})$, where S_i denotes the support of word-cluster i, a real-valued vector representing the textual factor F_i, and the factor importance d_i. Given a new data-point (document d) represented by a document-term vector $N^{(d)} \in \mathbb{R}^V$, the loadings of the textual factor i is simply

$$x_i^{(d)} := \frac{\langle N_{S_i}^{(d)}, F_i \rangle}{\langle F_i, F_i \rangle} \tag{1}$$

and the document D can be represented quantitatively as $(x_1^{(d)}, \dots x_K^{(d)}) \in \mathbb{R}^K$.

To understand the meaning of these loadings, take publicly listed firms for example. A company discloses numbers on revenues, profits, liabilities, etc. Texts about the company could also touch on profitability, social responsibility, innovativeness, etc., each of which is a topic. The $x_k^{(d)}$ we obtain allows us to assign a coefficient that measures how much the company is exposed to that topic—a metric we can obtain in simple sparse regressions.

Finally, the authors remark that one can easily generalize their methodology to apply to document-term matrix that include multi-grams. And in that case, one can significantly reduce the dimensionality of the multi-gram space by considering multi-grams with words in only one or say a few topics.

Illustrations. To check whether the textual factors generated make sense, the authors first use a few examples to illustrate their interpretability. They first compare the word clusters generated from google word embedding with the plain-vanilla LDA, and print-out the support of the word clusters. Table 1 displays the top three obtained "clusters", or topics by plain-vanilla LDA. As we can see, extremely common words dominate each cluster, which clouds the meaning of different topics. In contrast, Table 2 illustrates the effectiveness of their clustering method based on LSH. The gain in interpretability is apparent.

They also test the robustness and sensibility of loadings on their constructed textual factors. Specifically, they inspect the trends of loadings over time from 1900-2000 on *Wall Street Journal* article titles, for representative clusters such as "Recession", "War" and "Computer". From the plots of loadings over time, their results seem plausible because the intensity of the textual factors accurately captures the prominence of these topics in history.

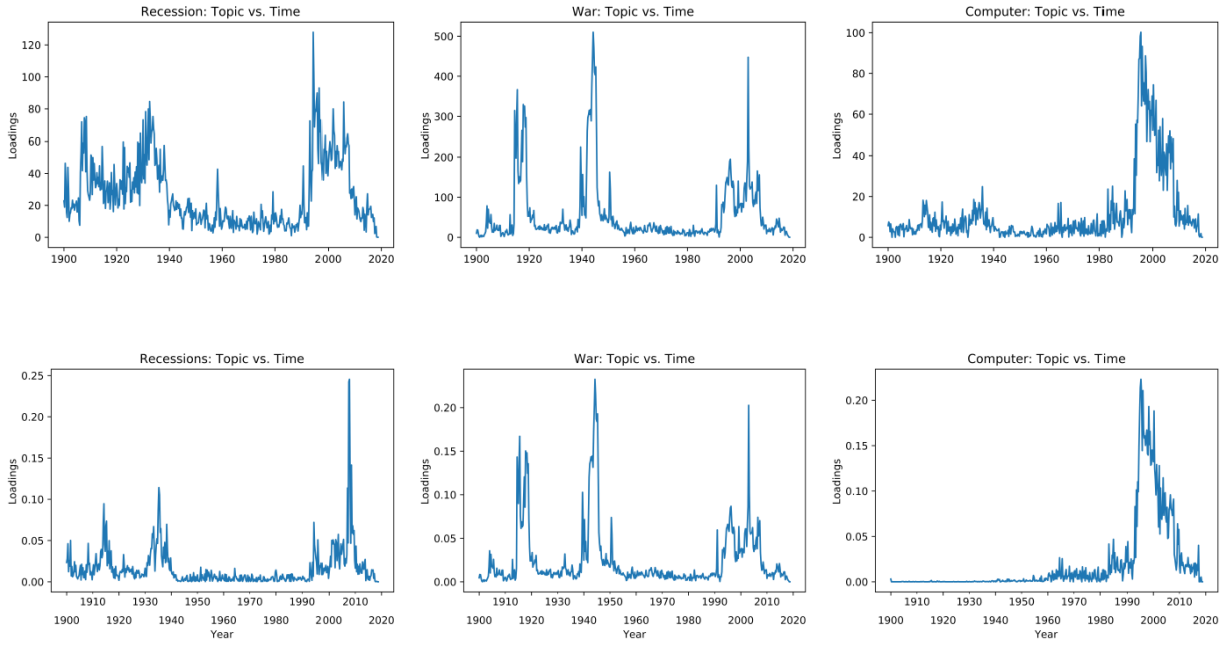

Figure 3: Loadings on textual factors over time, WSJ data. The three columns correspond to "Recession", "War" and "Computer", respectively. Reproduced from Cong, Liang, and Zhang (2019).

Cluster	Support
Topic ID: 62, Prob: 0.20071%	**washington**, tax, business, york, labor, letter, bulletin, wire, report, old, many, big, **president**, like, long, economic, **price**s, time, ago, federal, outlook, city, get, high, sales, white, house, back, people, even, state, just, home, world, much, **american**, man, next, **government**, job, million, still, work, companies, workers, economy, men, three, little
Topic ID: 1272, Prob: 0.17438%	stock, dividend, steel, business, **american**, oil, common, market, york, **earnings**, months, outlook, **cent**s, made, record, way, chicago, share, company, united, net, time, **president**, rate, **price**s, increase, railroad, states, june, **price**, **general**, review, shares, declared, july, report, cotton, preferred, sales, **washington**, present, large, month, regular, production, exchange, pacific, cars, quarterly, september
Topic ID: 1828: Prob: 0.11747%	steel, states, business, united, outlook, review, railroad, stock, way, market, york, country, time, **president**, great, made, **american**, **price**s, copper, increase, **earnings**, corporation, public, **government**, per, national, **general**, since, **washington**, cotton, crop, bank, report, months, state, much, commission, present, **cent**, railroads, rate, conditions, **price**, large, street, ago, letter, pacific, trade, three

Table 1: Sample plain-vanilla LDA clusters. Reproduced from Cong, Liang, and Zhang (2019).

Cluster	Support
Tax	quotas, visa, harvestable, import, preferential, abolished, tariffs, quota, sanction, compulsory, tariff, compulsorily, stipulating, fisheries, cess, exports, pricing, export, telcos, exporters, import, liberalization, preferential, excise, tariffs, tax, tariff, importers, deregulation, antidumping, subsidy
Oil	refiners, refiner, refineries, refinery, petrochemical, feedstock, refiners, pipelines, smelters, crudes, oil, bpd, gasoline, refiner, petrochemicals, petroleum, refining, ethanol, refineries, tankers, refinery, coker, petrochemical, ethylene, feedstock, crude
Unemployment	stimulus, foreclosures, recession, claimants, workweek, unemployed, housing, unemployment, jobless, economy, workers
Volatility	correction, uptrend, readjustment, reversal, retest, revision, divergence, retrenchment, steepening, selloff, rebalancing, bearish, pullbacks, corrective, correcting, reversion, stabilization, selldown, snapback, reassessment, volatility, pullback, bull, corrections, bottoming, downtrend
Exports	consignments, foodstuffs, exports, tins, cargo, goods, warehouses, equipments, importers, exporting, containers, tonnages, exporters, import, imports, perishable, cartons, cargoes, export, adulterated, tankers, pallets, wholesalers, demurrage, customs, transporters, consignment, consignee, exported
Investment	development, capitalization, differentiation, invest, macro, optionality, strategic, capex, macroeconomic, countercyclical, investments, investing, outperformance, diversification, equity, arbitrage, diversify, cyclicality, underperformance, diversifying, expansion, diversified, geographies, reinvest, specialization, profitability, deleveraging, consolidation, renewables, volatility, investment, liquidity, growth, maximization, sector, cyclical, synergy, reinvesting, investors, reinvestment
Stimulus	appropriation, moneys, underfunded, money, reauthorization, subsidies, budget, fundings, budgeted, allocations, budgets, budgetary, stimulus, funded, appropriations, funds, grant, nonfederal, appropriated, earmarked, infrastructure, reauthorized, assistance, unfunded, funding, financing, grants, monies, support, underfunding
Disasters	disturbances, occurance, instances, recur, disasters, incidences, occur, occurence, occurrences, causes, occurred, occurrence, phenomenon, earthquakes, anomaly, outbreaks, accidents, incidents, emergencies, observations, tragedies, ultramafic, catastrophes, polymetallic, anomalous, occurence, outbreaks, disturbances, incidences, occur, calamities, occurrences, infrequent, phenomena, anomalies, occurance, happening, intrusions, contaminations, occurrence, occurring, incidents
War	battles, confrontation, dispute, fighting, showdown, struggle, fight, battle, wars, fierce, war, battles, confrontation, showdown, matchups, fight, battle, victory
Election	political, intellectual, politically, election, politicians, democratic, religious, republican, incumbency, diplomatic, politics, economic

Table 2: Sample clusters reproduced from Cong, Liang, and Zhang (2019).

Applications. The authors describe three ways the textual factor framework can be applied. First, textual factors can help predict or explain outcomes in cross-section, time series, and panel data analysis. For example, one can use newspaper front page titles and abstracts to forecast macroeconomic outcomes such as CPI or to train a model to better understand factors driving market volatility or to backfill the VIX index in a way similar to Manela and Moreira (2017).

Second, they can be used to interpret existing explanatory variables constructed from structured data, such as Fama-French three factors, or patent citations. For example, discussions on risk factors or MD&A from company filings can provide useful information on the cross-sectional beta loadings on the Fama-French three factors.

Finally, they allow a data-driven way for constructing explanatory variables or metrics. This last dimension also points to the possibility for textual factors to create new domain knowledge and opens new frontiers of analysis. For example, using structured data such as revenue and user base to value start-ups has been challenging because most early projects do not generate stable or positive cash flow, and their valuation largely depends on investors' beliefs and perception. In contrast, information extracted from unstructured data in news, forum discussion, user feedback and ratings can provide meaningful insights into start-up's valuation. Another example is to use texts to construct metrics of market sentiments, building on earlier work by Garcia (2013). One can use proxy statements to measure corporate governance (Cong and Malenko, 2019) or patents and stock prices to measure innovation (e.g., Chen, Wu, and Yang, 2019).

In what follows, we take their methodology to backfill expectation errors in the credit market. This is an important issue because academics and policymakers debate over whether expectation errors predict future macroeconomic outcomes. To answer this question, we need a long time series of expectation data. However, forecast data of credit spread was available only for recent years. To conquer this problem, we backfill expectation error data by applying textual analysis to article titles in the WSJ. We use Blue Chip Financial Forecast data from

1999 to 2017 as our training sample and use text data to backfill expectation from 1929 to 1998.

To backfill expectation error, we first estimated the following model:

$$error_t = \alpha + \gamma x_t^T + \eta_t, \tag{2}$$

where $error_t$ is the difference between the expectation and the realized Baa corporate bond spread. The expectation of Baa corporate bond spread is defined as consensus forecast of Baa corporate bond yield minus consensus forecast of 10-year Treasury yield. Both are one-year forecasts collected from Blue Chip Financial Forecasts. The realized Baa corporate bond Spread is calculated in the same way using historical value.

To further manage model dimensionality, we apply LASSO penalization to estimate (2). We find that discussions about government (e.g. taxes, president, white house, and Washington), finance (money, banks, treasury, credit, and stock), recessions (e.g. great depressions, great recessions, crisis, and economic downturns), war (e.g. military, world war, and Iraq) are the most useful in constructing expectation error. Using estimated γ and topic loadings, we backcast expectation errors for a long horizon, as shown in Figure 4.

A clear pattern emerges from Figure 4: expectation error tends to be positive (overly optimistic) at the end of booms and negative (overly pessimistic) during recessions. The countercyclical nature suggests that expectation error may predict business cycles. We explore this pattern more carefully in the following predictive regression framework:

$$\Delta y_{t+h} = \beta_0 + \beta_1 \widehat{error}_t + \sum \beta_j controls_{j,t} + \epsilon_{t+h}$$

where Δy_{t+h} is the log-difference of real GDP per capita over the course of year t + h. $\widehat{error}_t$ is the backfilled expectation error averaged over year t-1 to year t. $controls_{j,t}$ includes change in credit spreads over year t, change in GDP per capita from year t-1 to t, CPI inflation rate,

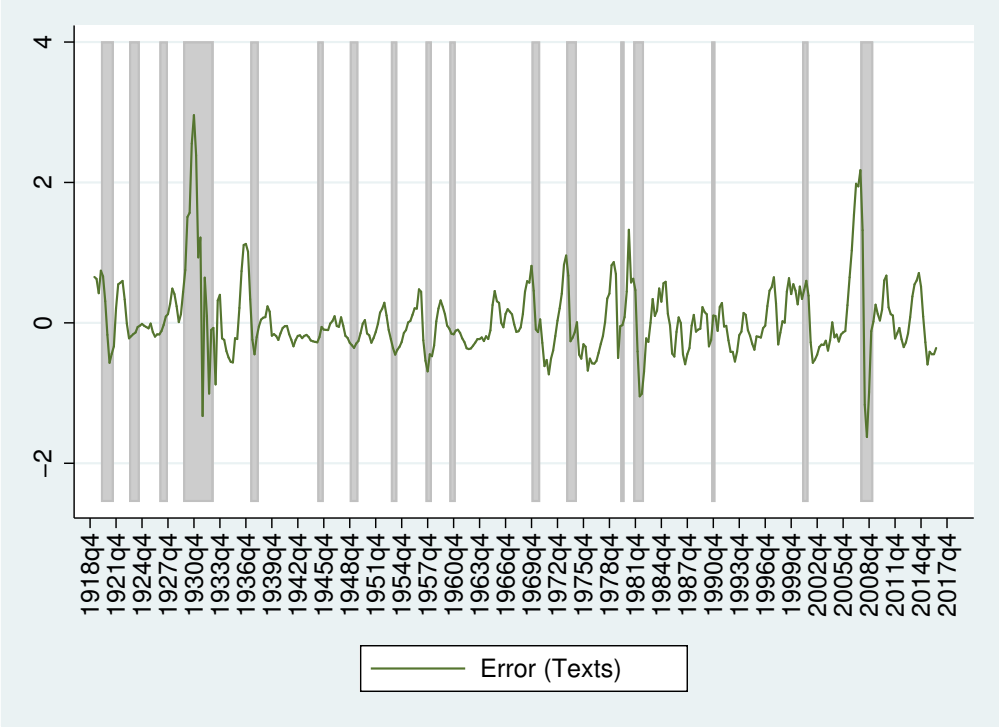

Figure 4: Backfilling Expectation Error

and changes in short-term and long-term Treasury yields. As a robustness check, we also include several lags of the control variables to ensure that mean-reversion in GDP growth is not responsible for the results.

Table 4 presents various specification of the predictive regression for different horizon. The explanatory variable of interest in this table is $\widehat{error}_t$. From Columns 1 to 3, we vary one-year output growth on the left-hand side from being contemporaneous to two years into the future. As can be seen from Column 2, expectation error at t have substantial forecasting power for GDP growth in year t+1 and t+2, even after controlling for changes in credit spread: a one standard deviation increase in expectation error is associated with a step-down in real GDP growth per capita of 0.45-0.5 standard deviations, or about 1.2 percentage points. In Columns 4 to 6, we add levels of credit spread as an additional control. The results remain largely unchanged. Neither changes nor levels of credit spread are predictive of real GDP growth in year t+1 or t+2. Instead, expectation error is a strong

Table 4:Predictive Regressions: Real GDP Growth

	(1) h=0	(2) h=1	(3) h=2	(4) h=0	(5) h=1	(6) h=2
$\Delta ExpectationError_{t-1}$	0.004	-0.022***	-0.021***	0.005	-0.020***	-0.021***
	(0.005)	(0.006)	(0.006)	(0.005)	(0.007)	(0.006)
$\Delta CreditSpread_{t-1}$	-0.043***	0.004	0.004	-0.044***	-0.001	0.002
	(0.008)	(0.010)	(0.010)	(0.008)	(0.011)	(0.010)
$CreditSpread_{t-1}$				0.001	0.007	0.002
				(0.005)	(0.006)	(0.004)
R^2	0.552	0.268	0.212	0.552	0.287	0.216

$control_{j,t}$ also include change in GDP, and other significant variables documented in literature such as CPI inflation rate and changes in short-term and long-term Treasury yields. ***,**,* indicate coefficient estimates statistically different than zero at the 1%, 5% and 10% confidence level, respectively.

predictor of future GDP growth.

7 Other Approaches and Promising Directions

The textual-factor approach is just one of the many plausible ways to improve textual analysis in social sciences. While there have been several attempts over the past few years, a few directions are especially worth highlighting. Instead of providing an exhaustive list, we discuss two of them.

7.1 Dynamic & Customized Count-based Methods

The count-based approach can be further extended to analyze questions economists care about. For example, Hoberg and Phillips (2016) develop a new time-varying measurement of product similarity using business descriptions in 10-K filings to compute pairwise word similarity scores for each pair of firms in a given year. Specifically, they represent each text document using a vector, with each element being populated by the number 1 if that text uses the given word and 0 if it does not. Then they calculate the firm pairwise similarity

score using cosine similarity formula. They find that their measure of product similarity is much better than the traditional ways such as using SIC and NAICS classification code, because their measure allows industry competition to be firm-centric and change over time. Equipped with this new measure, they study questions related to theories of endogenous product differentiation and find that firm RD and advertising are associated with subsequent differentiation from competitors. Using a similar methodology, Hoberg and Phillips (2010) find that firms with more similar product descriptions are more likely to enter mergers agreements and experience increased stock returns and real longer-term gains in cash flows and higher growth.

While having the appearance of being count-based, they are not susceptible to the usual limitations of count-based methods because they are dynamic and customized. To see this, Hoberg and Phillips (2016) uses no pre-determination of vocabularies. The dictionaries are instead dynamic and customized to each firm based on general economic foundations regarding the concept of competition and rivalry. Specifically, each document is being scored to a different (dynamically selected) set of documents for comparison.[6]

Such dynamism and customization also manifest in Hanley and Hoberg (2010) in which the dictionary is customized both in time and by industry, and in some cases by underwriter. One interesting sub-category of the dynamic count-based models comprises of studies looking at "document revision intensity." Brown and Tucker (2011) is an important early study in ACCT using MDAs. Hanley and Hoberg (2012) use IPO prospectuses and finds that a lack of content revision, when there is a price revision, indicates a situation where litigation and high underpricing are likely. More recently, Cohen, Malloy, and Nguyen (2018) show that active change in firms' reporting practices conveys an important signal about future firm operations and affect share prices.

In all studies, the authors customize comparisons and utilized dynamic word lists based

[6]For example, a given firm in a given year is scored relative to the set of other firms in its neighborhood spatially and its own unique business description. This comparison is different for every firm without fixed global word lists, and also changes over time as the documents evolve.

on economic principles. Such a dynamic/customized extension to cound-based methods are rather generalizable and flexible. For example, document revision has potential in many more settings. For extracting and analyzing information from textual data within a reasonable size and time frame, the dynamic/customized count-based approach holds great promise.

7.2 Machine Learning for Economics

Many machine learning tools often has black-box appearances and are hard to understand or interpret. Applying them to textual analysis without an effort to understand the underlying mechanism or economic content is likely futile. After all, we are economists aiming to contribute to philosophical and economic knowledge, not just raw predictability without insight.

That said, many recent developments in natural language processing, once properly used, provide supplementary data structures that are extremely informative about what drives any given signal. For example, topic models such LDA and LSA generate factors with their word lists, and word2vec gives an entire embedding matrix with a representation of each word that can be used to illustrate the content that resulted in any outcome. The key is to develop analytics using them in a transparent and interpretable manner. The textual-factor framework we introduced in the previous section is an example in this direction.

Another example is Hanley and Hoberg (2019) who combine a LDA model and word2vect (referred to as semantic vector analysis in their paper) to identify emerging risks using bank 10-K risk factor section. They apply LDA model to identify topics that are important in explaining time-series variation in risk that banks face. In their research setting, this approach is better than count-based methods that require domain knowledge because the sources of financial instability are inherently unpredictable and might be unknown ex ante to the researchers. Using the most representative words associated with each topic, they construct risk exposure to different emerging risk by calculating cosine similarity based on semantic vectors. They find that the two models in tandem does a good job in detecting

emerging risks. In addition, the elevated risks predict the financial crisis of 2008 well before VIX or aggregate volatility does. At the individual bank level, they find that banks with greater ex-ante exposure to emerging risks experience significantly lower stock returns during the financial crisis, lending credibility to their methodology.

Both Hanley and Hoberg (2019) and Cong, Liang, and Zhang (2019) derive factor structures from texts. While Cong, Liang, and Zhang (2019) aims at a general factor-generation tool for textual analysis, Hanley and Hoberg (2019) has the goal of detecting systemically important risk factors pervasive across many banks that are interpretable so financial instability can be detected early, before linking them to the covariance matrix of bank stock prices. This could facilitate pre-emptive research by regulators and potentially pre-emptive policy remedies that can reduce damage before instability becomes crisis. Their orders of applying word2vec and topic modeling are also different. The two papers complement each other well and the efficacy of their approaches jointly underscore the extraordinary value of utilizing embedding and NLP tools actively developed in computer science and statistics for applications and methodologies in economics and social science.

8 Concluding Remarks

Modern institutions leverage big data for originating loans, predicting asset returns, improving customer service, etc. Texts, as a form of unstructured data, are abundant and their interpretability sheds light on key economic mechanisms and explanatory variables. We discuss recent developments in textual analysis and its applications in finance and economics. We highlight the need for a framework for analyzing large-scale text-based data that can capture complex linguistic structures while ensuring computational scalability and economic interpretability. As Athey (2018) predicts, "extensions and modifications of prediction methods to account for considerations such as fairness, manipulability, and interpretability to be among the very first changes to emerge concerning how empirical work is conducted." A few approaches combining the strengths of neural network language models and generative

statistical modeling aim to balance model complexity and interpretability, which may prove to be promising directions and useful analytic tools for future research.

References

Andoni, Alexandr, Piotr Indyk, Thijs Laarhoven, Ilya Razenshteyn, and Ludwig Schmidt, 2015, Practical and optimal lsh for angular distance, in *Advances in Neural Information Processing Systems* pp. 1225–1233.

Antweiler, Werner, and Murray Z Frank, 2004, Is all that talk just noise? the information content of internet stock message boards, *The Journal of finance* 59, 1259–1294.

Athey, Susan, 2018, The impact of machine learning on economics, in *The economics of artificial intelligence: An agenda* (University of Chicago Press).

Baker, Scott R, Nicholas Bloom, and Steven J Davis, 2016, Measuring economic policy uncertainty, *The Quarterly Journal of Economics* 131, 1593–1636.

Bandiera, Oriana, Renata Lemos, Andrea Prat, and Raffaella Sadun, 2017, Managing the family firm: evidence from ceos at work, *The Review of Financial Studies* 31, 1605–1653.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin, 2003, A neural probabilistic language model, *Journal of machine learning research* 3, 1137–1155.

Bollen, Johan, Huina Mao, and Xiaojun Zeng, 2011, Twitter mood predicts the stock market, *Journal of computational science* 2, 1–8.

Brown, Stephen V, and Jennifer Wu Tucker, 2011, Large-sample evidence on firms year-over-year md&a modifications, *Journal of Accounting Research* 49, 309–346.

Chen, Mark A, Qinxi Wu, and Baozhong Yang, 2019, How valuable is fintech innovation?, *The Review of Financial Studies* 32, 2062–2106.

Cohen, Lauren, Christopher Malloy, and Quoc Nguyen, 2018, Lazy prices, Discussion paper, National Bureau of Economic Research.

Cong, Lin William, Tengyuan Liang, and Xiao Zhang, 2019, Textual factors: A scalable, interpretable, and data-driven approach to analyzing unstructured information, *Working Paper*.

Cong, William, and Nadya Malenko, 2019, A textual factor approach to measuring corporate governance, *Work in Progress*.

Das, Sanjiv Ranjan, et al., 2014, Text and context: Language analytics in finance, *Foundations and Trends® in Finance* 8, 145–261.

Datar, Mayur, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni, 2004, Locality-sensitive hashing scheme based on p-stable distributions, in *Proceedings of the twentieth annual symposium on Computational geometry* pp. 253–262. ACM.

Evans, James A, and Pedro Aceves, 2016, Machine translation: mining text for social theory, *Annual Review of Sociology* 42.

Garcia, Diego, 2013, Sentiment during recessions, *The Journal of Finance* 68, 1267–1300.

Gentzkow, Matthew, Bryan T Kelly, and Matt Taddy, 2017, Text as data, Discussion paper, National Bureau of Economic Research.

Gentzkow, Matthew, Jesse Shapiro, Matt Taddy, et al., 2016, Measuring polarization in high-dimensional data: Method and application to congressional speech, Discussion paper, .

Grimmer, Justin, and Brandon M Stewart, 2013, Text as data: The promise and pitfalls of automatic content analysis methods for political texts, *Political analysis* 21, 267–297.

Hanley, Kathleen Weiss, and Gerard Hoberg, 2010, The information content of ipo prospectuses, *The Review of Financial Studies* 23, 2821–2864.

——— , 2012, Litigation risk, strategic disclosure and the underpricing of initial public offerings, *Journal of Financial Economics* 103, 235–254.

——— , 2019, Dynamic interpretation of emerging risks in the financial sector, *Review of Financial Studies* Forthcoming.

Hansen, Stephen, Michael McMahon, and Andrea Prat, 2017, Transparency and deliberation within the fomc: a computational linguistics approach, *The Quarterly Journal of Economics* 133, 801–870.

Hassan, Tarek A, Stephan Hollander, Laurence van Lent, and Ahmed Tahoun, 2017, Firm-level political risk: Measurement and effects, Working Paper 24029 National Bureau of Economic Research.

Heckman, James J, 1979, Sample selection bias as a specification error, *Econometrica* 47, 153–161.

Hoberg, Gerard, and Gordon Phillips, 2010, Product market synergies and competition in mergers and acquisitions: A text-based analysis, *The Review of Financial Studies* 23, 3773–3811.

——— , 2016, Text-based network industries and endogenous product differentiation, *Journal of Political Economy* 124, 1423–1465.

Huang, Allen H, Reuven Lehavy, Amy Y Zang, and Rong Zheng, 2017, Analyst information discovery and interpretation roles: A topic modeling approach, *Management Science* 64, 2833–2855.

Jegadeesh, Narasimhan, and Di Andrew Wu, 2017, Deciphering fedspeak: The information content of fomc meetings, .

Kearney, Colm, and Sha Liu, 2014, Textual sentiment in finance: A survey of methods and models, *International Review of Financial Analysis* 33, 171–185.

Kelly, Brian, Asaf Manela, and Alan Moreira, 2018, Text selection, *Working Paper*.

Li, Feng, 2010, Survey of the literature, *Journal of accounting literature* 29, 143–165.

Li, Kai, Feng Mai, Rui Shen, and Xinyan Yan, 2018, Measuring corporate culture using machine learning, *Available at SSRN 3256608*.

Loughran, Tim, and Bill McDonald, 2011, When is a liability not a liability? textual analysis, dictionaries, and 10-ks, *The Journal of Finance* 66, 35–65.

——— , 2016, Textual analysis in accounting and finance: A survey, *Journal of Accounting Research* 54, 1187–1230.

Manela, Asaf, and Alan Moreira, 2017, News implied volatility and disaster concerns, *Journal of Financial Economics* 123, 137–162.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, 2013, Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, 2013, Distributed representations of words and phrases and their compositionality, in *Advances in neural information processing systems* pp. 3111–3119.

Nini, Greg, David C Smith, and Amir Sufi, 2012, Creditor control rights, corporate governance, and firm value, *The Review of Financial Studies* 25, 1713–1761.

Pennington, Jeffrey, Richard Socher, and Christopher Manning, 2014, Glove: Global vectors for word representation, in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* pp. 1532–1543.

Sontag, David, and Dan Roy, 2011, Complexity of inference in latent dirichlet allocation, in *Advances in neural information processing systems* pp. 1008–1016.

Taddy, Matt, 2015, Document classification by inversion of distributed language representations, *arXiv preprint arXiv:1504.07295*.

Tetlock, Paul C, 2007, Giving content to investor sentiment: The role of media in the stock market, *The Journal of finance* 62, 1139–1168.

Wallach, Hanna M, David M Mimno, and Andrew McCallum, 2009, Rethinking lda: Why priors matter, in *Advances in neural information processing systems* pp. 1973–1981.

Wisniewski, Tomasz Piotr, and Brendan Lambe, 2013, The role of media in the credit crunch: The case of the banking sector, *Journal of Economic Behavior & Organization* 85, 163–175.