Text Algorithms in Economics*

Elliott Ash
ETH Zurich

Stephen Hansen
Imperial College London and CEPR

August 16, 2022

Abstract

This paper provides an overview of the methods used for algorithmic text analysis in economics, with a focus on three key contributions. First, the paper introduces methods for representing documents as high-dimensional count vectors over vocabulary terms, for representing words as vectors, and for representing word sequences as embedding vectors. Second, the paper defines four core empirical tasks that encompass most text-as-data research in economics, and enumerates the various approaches that have been taken so far for these tasks. Finally, the paper flags limitations in the current literature, with a focus on the challenge of validating algorithmic output.

*Ash and Hansen gratefully and respectively acknowledge financial support from ERC Starting Grant 101042554 and ERC Consolidator Grant 864863. For helpful feedback on earlier drafts, we thank Afra Amini, Sascha Becker, Nick Bloom, Prashant Garg, Bryan Kelly, Asaf Manela, Claudia Marangon, Hannes Mueller, Yabra Muvdi, Carlo Schwarz, Dominik Stammbach, Szymon Sacher, Bryan Seegmiller, Francesca Truffa, and Ashley Wong. Laura Battaglia, Claudia Marangon, and Yabra Muvdi provided excellent research assistance. We thank GPT-3 for writing the abstract.

1 Introduction

Text analysis in economics is not new. Classic examples include Coase (1960), who analyzed legal cases to study how the law resolves externality problems, and Friedman and Schwartz (1963), who pioneered the construction of policy surprises via historical documents. Until recently, though, text analysis was conducted via careful human reading, which cannot be scaled to the massive corpora now available.[1] The number of individual documents in popular databases such as newspaper and job posting records can easily add up to tens of millions. The increasing availability of large-scale corpora has led to increasing interest in algorithmic text analysis, a trend which will likely continue to grow as more text data becomes available.

Because text algorithms are relatively new to economists, there is little consensus on how best to deploy them. There is substantial methodological diversity and no common framework, nor even vocabulary, for understanding what should guide modeling choices. This difficulty is reinforced by the rapid pace of development of natural language processing (NLP) algorithms: even since Gentzkow et al.'s (2019) review of text-as-data methods in economics, NLP has been revolutionized by a new generation of deep neural network models, known as Transformers, that can detect subtle patterns and semantic meaning in language.

We cannot do justice to the vast NLP literature nor to the varied text-as-data applications in economics and other social sciences. In our limited space, therefore, we focus on three contributions. First, Section 2 provides a conceptual overview of the methods that now form the basic building blocks of algorithmic text analysis in economics. We start with methods that represent documents as high-dimensional count vectors over vocabulary terms and reduce their dimensionality with latent factor models. Next, we review methods for representing words as vectors (also known as 'word embeddings'), constructed using information on local co-occurrence patterns such that words with similar meanings have proximate vectors. Lastly, we introduce recently developed methods for representing word *sequences* as embedding vectors. These sequence models allow relationships among words to inform meaning—for example, while word embedding models assign a fixed vector to '*class*', sequence embedding models allow meaning to depend on neighboring words, with distinct vectors assigned for 'she filed suit under *class* action' and 'she graduated top of *class*'.

To illustrate the implementation and application of these algorithms, we build and refer to a companion GitHub page (`github.com/sekhansen/text_algorithms_econ`) with reusable code for teaching and research purposes. The examples use publicly available

[1] The explosion of information associated with the 'big data' revolution has mostly been driven by a growth in unstructured text, which now constitutes a large majority of the data in the world (Rydning 2021).

data and source code so that readers can replicate our results and extend them to other datasets. This repository will be regularly updated as new algorithms are introduced in the literature.

Our second contribution, described in Section 3, is to define four core measurement problems that encompass most text-as-data research in economics: 1) measuring the similarity among documents; 2) measuring economic concepts contained in raw text; 3) measuring how concepts are related to each other in text; and 4) relating text to quantitative metadata. Even as algorithms develop and change, they will mostly add value to economics insofar as they help solve one of these problems. We enumerate the various approaches the literature has taken so far for these problems and assess the relevant trade-offs to design choices. Section 4 then briefly discusses econometric issues that arise from using these measures in downstream regression models.

Finally, in Section 5, we flag limitations in the current literature. Perhaps most important is the challenge of validating algorithmic output. Economists generally have a different notion of relevance than computer scientists, so merely borrowing validation tasks from other fields is insufficient. To illustrate the problem, we adapt eleven popular algorithms for computing document similarity and apply each of them to a corpus of firms' annual regulatory filings. The different similarity measures frequently disagree on which documents are most similar, which in turn affects inference in downstream regression models associating textual similarity with firm covariates. Deciding which similarity metric is preferred therefore requires human judgment.

More generally, our view is that the traditional mode of text analysis involving human readers with domain expertise and the algorithmic mode are clear complements, and the latter should not replace the former. It is natural that, in the early stages of adoption, more attention is placed on the introduction of new algorithms and measures than on evaluating their performance against a common benchmark informed by human judgment. As the literature matures, though, we expect it to move in this direction. We hope readers of this review will be inspired to hasten the process.

2 Text Algorithms

Our discussion of algorithms lays out the main ideas and motivations. Implementation details are best illustrated through code, which we provide at `https://github.com/sekhansen/text_algorithms_econ`. The important set of questions around specifying model parameters are part of the larger issue of setting criteria for on algorithm choice, which we delay to Section 5.

2.1 Preliminaries

Algorithmic text analysis starts with a machine-readable collection of D documents. Reaching this point can be a challenge in itself, as text data may only be available embedded in a markup language (e.g. HTML or XML) or in scanned image files (e.g. PDFs of historical books). In these cases, relevant text and metadata must be extracted and organized before any analysis begins.[2]

In turn, a relevant design decision is how to define a "document." For example, when using financial newspaper articles for macroeconomic forecasting, one might aggregate all articles together at the relevant time frequency (daily, quarterly, etc). Another consideration is that the performance of algorithms can depend on document length. Linguistic parsing algorithms for determining grammatical structure typically operate at the sentence level, for example, while modern attention-based neural network models have a limit on the length of document inputs.

Before one applies any algorithm, raw document text must be converted into sequences of linguistic features, called *tokens* in the NLP literature. We denote the content of document d as $\mathbf{w}_d = (w_{d,1}, \ldots, w_{d,t}, \ldots, w_{d,N_d})$, where the encoded features used as tokens, and the sequence of *preprocessing* operations to obtain $\{\mathbf{w}_d\}_{d=1}^{D}$, will vary across applications. In economics, the standard preprocessing approach is to represent documents as lists of words, typically reduced to some root form. The standard approach has been extensively discussed in other material (e.g. Manning et al. 2008, Grimmer and Stewart 2013, Denny and Spirling 2018a, Gentzkow et al. 2019a). The basic steps are tokenizing (splitting on whitespace/punctuation), dropping non-letter characters, dropping common *stopwords* like 'the'/'to'/'is', adjusting letters to lowercase, and stemming words to remove suffixes.[3] It is also standard to capture information on local word order by producing n-grams — phrases up to length n — from these pre-processed word lists. The resulting elements of $\mathbf{w}_d$ are often called *terms* and in general are no longer properly spelled English words.

There are two other pre-processing approaches bearing mention which so far are less used in economics. For some applications, it is useful to add additional grammatical information on the functions of and relations between words, using linguistic annotation

[2]Usually researchers rely on existing software packages for HTML/XML parsing (e.g. `Beautiful Soup` in Python) and optical character recognition (e.g. `Layout Parser` in Python), then use regular expressions to further clean and organize the output. Machine-learning-based data segmenting is often not worth the decrease in transparency, but these algorithms are improving rapidly and can be helpful for separating documents on the same page, for example articles in historical newspapers (e.g. Shen et al. 2021).

[3]Stemming consolidates grammatically distinct but conceptually identical words like 'walked' and 'walking' into a single stem 'walk'. The Porter stemmer is a common default. This need not output English words, and in some cases incorrectly consolidates words. An example is 'university' and 'universe' whose stems are 'univers'. An alternative is to instead *lemmatize* words by searching for linguistic roots in a dictionary. See the companion GitHub repo for additional details on preprocessing.

algorithms (e.g. Jurafsky and Martin 2020). In state-of-the-art neural network models of language, meanwhile, the standard approach is designed to neither add nor remove information—that is, to split plain texts into tokens without changing the text.[4]

2.2 Bag-of-words model

One popular representation of documents is the *bag-of-words* model. This begins by assigning to each unique vocabulary term a unique index value from the integers $1, \ldots, V$ where V is the number of unique terms.[5] Let $x_{d,v} = \sum_n \mathbb{1}(w_{d,n} = v)$ be the count of term v in document d, $\mathbf{x}_d = (x_{d,1}, \ldots, x_{d,V})$ be the vector of counts, and X the *document-term matrix* formed by stacking the $\mathbf{x}_d$ across rows.

The data representation X forms the core of much of text analysis in economics. Two properties of X distinguish it from the usual matrix-structured dataset. First, it has a vast number of columns: even in small corpora, V can be on the order of tens of thousands. Second, X is sparse, since most vocabulary terms v are not present in the average document, i.e. $x_{d,v} = 0$ for almost all $v \in \{1, ..., V\}$.

2.3 Dimensionality reduction

Often we care not about the particular words documents use, but about the underlying meaning those words reflect. The documents 'investors fear rising prices' and 'market participants are anxious about inflation' share no common terms, so their bag-of-words representations would produce orthogonal term-count vectors. Yet they clearly have the same meaning. In the same way that factor analysis is designed to capture structure in high-dimensional economic data, dimensionality reduction in NLP can be viewed as projecting documents into a 'meaning' space which reflects more relevant heterogeneity than the high-dimensional term space.

For economists, one of the most familiar dimensionality reduction techniques is principal components analysis (PCA). Applying PCA to the document-term matrix is known as **latent semantic analysis** (LSA) and is one of the earliest dimensionality reductions used in text analysis (Deerwester et al. 1990). The $K < V$ resulting principal components represent latent thematic content recovered from document-level term co-occurrence patterns, in the same way that principal components produced from high-dimensional economic datasets represent deeper structure.

[4] To reduce the number of characters, capitalization is represented by a special prefix token before a lowercase letter. Words are broken into separate pieces (e.g. 'walking' becomes 'walk' and 'ing'), to help neural models learn more meaningful word representations, especially for rare and long words. This is called byte-pair encoding (e.g. Goldberg 2017).

[5] The choice of which terms are assigned to which indices is arbitrary, but it is often convenient to sort by corpus frequency.

While intuitively related to familiar techniques and straightforward to implement, LSA has unclear statistical foundations which can hinder interpretation of its outputs. The statistics literature has linked PCA and factor models for Gaussian distributions (Tipping and Bishop 1999), but $\mathbf{x}_d$ is discrete and sparse. An alternative approach to reducing the dimensionality of $\mathbf{x}_d$ begins with an explicit generative model of text, most commonly using the multinomial distribution:

$$\mathbf{x}_d \sim \text{Multinom}(\mathbf{q}_d, N_d). \tag{1}$$

In **probabilistic latent sematic analysis** (pLSA, Hofmann 1999), a factor model for discrete data is obtained by assuming that $\mathbf{q}_d = \sum_{k=1}^{K} \theta_{d,k}\boldsymbol{\beta}_k$. That is, documents are built from K common factors—or *topics*—each represented by a separate distribution over vocabulary terms $\boldsymbol{\beta}_k \in \Delta^{V-1}$. In turn, each document is characterized by a K-dimensional distribution over topics $\boldsymbol{\theta}_d \in \Delta^{K-1}$. pLSA thus reduces the dimensionality of documents from V to K like LSA, but within a more appropriate statistical model.

The likelihood function for pLSA is

$$\prod_d \prod_v \left(\sum_k \theta_{d,k}\beta_{k,v} \right)^{x_{d,v}} = \prod_d \prod_v \left[\left(\Theta B^T\right)_{(d,v)} \right]^{x_{d,v}} \tag{2}$$

where Θ is a $(D \times K)$ row-stochastic matrix and B is a $(V \times K)$ column-stochastic matrix. The right-side formulation suggests an alternative interpretation of pLSA based on matrix factorization (Ding et al. 2006). If we transform the term-count matrix X to a term-frequency matrix X'—i.e., divide each row d by the document length N_d—we can view maximization of (2) as finding the Θ and B that best approximate X'. That optimization objective is also known as **non-negative matrix factorization** (NMF).

In high-dimensional parameter spaces with sparse data, maximum likelihood estimation is prone to over-fitting. Moreover, the NMF of X' is not unique, so Θ and B are only set-identified (Ke et al. 2021). One solution to these problems is to place prior distributions over each probability vector $\boldsymbol{\theta}_d$ and $\boldsymbol{\beta}_k$ and use Bayesian inference for estimation. A computationally convenient choice of prior is the Dirichlet distribution—i.e. $\boldsymbol{\theta}_d \sim \text{Dir}(\alpha)$ and $\boldsymbol{\beta}_k \sim \text{Dir}(\eta)$—as the Dirichlet is conjugate to the categorical and multinomial distributions. Factoring X' with Dirichlet priors is known as **Latent Dirichlet allocation** (LDA) and has become ubiquitous in applications of text algorithms Blei et al. (2003).[6]

[6]The original LDA paper (Blei et al. 2003) only placed a Dirichlet prior on $\boldsymbol{\theta}_d$ terms and allowed it to be non-symmetric. Here we present the fully Bayesian LDA with symmetric priors, since this is most common in the economics literature. Typically η is chosen to be small to promote sparsity in the posterior distribution of the $\boldsymbol{\beta}_k$ vectors, in line with Zipf's Law approximately holding for term counts in natural language. Common defaults for α are 1, which imposes a uniform Dirichlet prior on $\boldsymbol{\theta}_d$, and $50/K$, as suggested by Griffiths and Steyvers (2004). See Wallach et al. (2009) for additional thoughts

LDA has gained popularity because it is computationally efficient and tends to produce human-interpretable topics more easily than other methods. Figure 1 illustrates the output of an estimated LDA model based on the transcripts of Federal Reserve Open Market Committee transcripts (Hansen et al. 2018). These topics intuitively indicate the importance of credit markets and other negative economic indicators during recessions.

2.4 Word embedding with local context

$\mathbf{x}_d$ represents documents as *global* counts over vocabulary terms independently of where they occur. However, semantic meaning is largely contained in the *local* context connecting words. While in principle the bag-of-words model can be extended locally by tabulating n-grams, in practice the feature space V quickly explodes in n. More subtly, a particular word's meaning may depend not just on immediate neighbors but on longer-range local dependencies within text.

An influential line of work in NLP reframes the analysis from statistics on document-term counts to statistics on each term's local co-occurrence with other terms. A family of algorithms known as *word embedding* models encodes and leverages the information in these local contexts for a panoply of NLP tasks. These models represent words as relatively low-dimensional and dense vectors, or *embeddings*, which compress the high-dimensional and sparse information on co-occurrence across the whole corpus.[7]

The embedding model **GloVe** ("Global Vectors") by Pennington et al. (2014) is explicitly designed to construct word vectors encoding local co-occurrence. Let the *context* of word $w_{d,n}$ be $C(w_{d,n}) = (w_{d,n-L}, \ldots, w_{d,n-1}, w_{d,n+1}, w_{d,n+L})$, i.e., a length-$2L$ window surrounding $w_{d,n}$. Then define W as the $V \times V$ word co-occurrence matrix, where an item W_{ij} is the number of times that word i appears within an L-tokens span of j (and vice versa, hence W is symmetric by construction). The choice of L depends on how one will use the resulting vectors, with shorter windows (e.g. $L = 2$) encoding more functional/syntactic word information, and longer windows encoding topics. With an arbitrarily large L, W_{ij} would count the number of times that word i co-occurs in the same document as word j. A standard parameter choice is $L = 10$.[8]

In GloVe, each vocabulary term v is associated with a word vector $\boldsymbol{\rho}_v$ in $\mathbb{R}^K$, with a standard parameter choice $K = 200$. These vectors are then chosen to solve

$$\min_{\boldsymbol{\rho}_v} \sum_{i,j} f(W_{i,j}) \left(\boldsymbol{\rho}_i^T \boldsymbol{\rho}_j - \log\left(W_{i,j}\right)\right)^2 \tag{3}$$

on prior selection in LDA.

[7]The term "embedding" comes from the neural network literature, in which an "embedding layer" is an input function that efficiently compresses high-dimensional data down to a low-dimensional dense representation for input to subsequent neural network layers.

[8]Such parameter choices are made through performance on standard evaluation tasks, such as solving analogies. Whether these tasks are relevant for economics is not clear.

where $f(\cdot)$ is a non-negative, increasing, and concave weighting function, such that rare word pairs count less in the objective.[9] Intuitively, GloVe's least-squares objective minimizes the squared difference between the dot product of the word vectors, $\boldsymbol{\rho}_i^T \boldsymbol{\rho}_j$, and the empirical co-occurrence, $\log(W_{ij})$. Terms that tend to co-occur are pushed in directions such that they have a high correlation (dot product).

An equally influential word embedding model is **Word2Vec** (Mikolov et al. 2013a,b), which treats each instance of a word and its context as a separate prediction problem that word vectors are chosen to solve. In addition to the word vector $\boldsymbol{\rho}_v$, each vocabulary term v is assigned a context vector $\boldsymbol{\alpha}_v$, also in the K-dimensional real numbers. Word2Vec parametrizes the probability of a word given its context as[10]

$$\Pr\left[\, w_{d,n} = v \mid C(w_{d,n}) \,\right] = \frac{\exp(\overline{\boldsymbol{\alpha}}_{d,n}^T \boldsymbol{\rho}_v)}{\sum_{v'} \exp(\overline{\boldsymbol{\alpha}}_{d,n}^T \boldsymbol{\rho}_{v'})} \text{ where } \overline{\boldsymbol{\alpha}}_{d,n} = \frac{1}{2L} \sum_{w \in C(w_{d,n})} \boldsymbol{\alpha}_w \tag{4}$$

Word2Vec learns word vectors and context vectors to maximize predictive accuracy of this model across all terms in the corpus.[11] In this sense, Word2Vec converts an unsupervised learning problem—finding latent dimensions of meaning in a large corpus—into a supervised learning problem, where the prediction target emerges from the structure of the corpus. Using prediction targets arising from language in the absence of external labels is known as *self-supervised learning*. The hope is that solving these auxiliary prediction problems with low-dimensional word vectors is informative about the latent meaning dimensions of primary interest.

With both GloVe and Word2Vec, the fitted word vectors $\hat{\boldsymbol{\rho}}_v$ are known as *embeddings*. Intuitively, these embedding algorithms give similar representations to words that appear in similar corpus contexts. These vectors can be used to represent and compare vocabulary terms, or in further downstream tasks as described below.[12]

A final point concerns the corpus used for embedding estimation. In an ideal world, a researcher would have a corpus large enough to estimate bespoke embeddings to capture word meanings specific to the application. With smaller datasets, though, there is not

[9] The standard function, from Pennington et al. (2014), is $f(x) = (x/x_{max})^\alpha$ for $x < x_{max}$ and $f(x) = 1$ otherwise, with $x_{max} = 100$ and $\alpha = 3/4$.

[10] This Word2Vec variant is called the *continuous bag-of-words* model. Another variant—the *skip gram* model—predicts $C(w_{d,n})$ given $w_{d,n}$.

[11] Formally, this is a one-layer neural network with softmax activation function. Direct maximization is prohibitively costly to implement, primarily because of the large number of probabilities (V) that need to be estimated per word. Instead, Word2Vec employs computational simplifications that approximate likelihood maximization.

[12] Further, it is instructive to compare GloVe and Word2Vec with the dimensionality reduction algorithms for the bag-of-words model. LSA, NMF, and LDA can also be viewed as producing word embeddings. In particular, the $(V \times K)$ matrix B from (2) contains a series of row vectors corresponding to each term in the vocabulary (see also Levy and Goldberg 2014). Those vectors contain information about word co-occurrence at the document level, rather than within a local context.

enough information to learn reliable vectors. In these cases, one can use *pre-trained* embeddings estimated on a large, auxiliary corpus and port them to the new application, a strategy known as *transfer learning*. A popular choice is to use embeddings estimated on generic English text like Wikipedia. While this approach is still relatively under-explored in economics, an issue with transfer learning is that it may not produce the most useful word representations for economics tasks. There could be gains from using more field-specific corpora for transfer learning.[13]

2.5 Embedding Sequences with Attention

Consider the following sentences, where [MASK] refers to an omitted word:

'As a leading firm in the [MASK] sector, we hire highly skilled software engineers.'

'As a leading firm in the [MASK] sector, we hire highly skilled petroleum engineers.'

Most people would predict that the omitted word in the first sentences relates to information technology, while in the second sentence it relates to energy. The key words for informing this inference are 'software' and 'petroleum', respectively, whereas words like 'hire' and 'leading' are less informative. Humans intuitively know how to adjust attention to words for prediction, yet GloVe and Word2Vec weight all words in the context window equally when constructing embeddings. A major breakthrough in NLP has been to train algorithms to also "pay attention" to the relevant features for prediction problems in a context-specific manner (e.g. Bahdanau et al. 2015, Vaswani et al. 2017).[14]

This idea is formalized with a *self-attention function*, which takes as input a sequence of initial token embeddings and outputs a sequence of new token embeddings that allow the initial embeddings to interact. Let $(\rho^0_{d,1}, \ldots, \rho^0_{d,N_d})$ be the initial embeddings that make up a document. The new embedding at each position n is given by

$$\rho^1_{d,n} = \sum_{n'=1}^{N_d} w_{(d,n),n'} \rho^0_{d,n'} \text{ where } \sum_{n'=1}^{N_d} w_{(d,n),n'} = 1. \tag{5}$$

That is, each embedding in the transformed sequence is itself a weighted average of the embeddings in the initial sequence. The non-negative attention weights $w_{(d,n),n'}$, which are estimated during model training, determine which pairs of (potentially distant) tokens interact to form each context-sensitive word embedding in the final document representation. The attention weights are estimated by a neural network to form embeddings

[13] For an illustration of this strategy, see Hansen et al. (2021).

[14] Besides neighboring words, another source of additional semantic information is the letters in the word itself. Bojanowski et al. (2017) provide a word embedding algorithm that constructs vectors from the constituent letters. This algorithm is especially useful for rare or unseen words, for example because they are misspelled due to OCR errors.

that are successful at whatever task the model is intended to solve, such as masked word prediction or other language-related prediction tasks.

Figure 2 shows an example of estimated attention weights for the masked word prediction task associated with the example above. As expected, the masked word interacts most strongly with 'software' and is most likely to be 'IT'. Heuristically, the model learns that references to certain occupations (software engineers) occur in the same postings as references to certain sectors (technology/automobile/health). In the masked word prediction problem, this is the important information and irrelevant tokens are ignored.

Besides these gains in capturing contextual semantic information, a major advantage of attention functions is that massive neural networks composed of stacked attention layers—known as **Transformers**—can be efficiently parallelized for training using specialized processors.[15] Attention is a core development behind the latest generation of Transformer-based, pre-trained language models, such as **BERT**/**RoBERTa** (Devlin et al. 2019, Liu et al. 2019) and **GPT** (Radford et al. 2018, Brown et al. 2020). Applying a self-supervised approach like Word2Vec, these models are pre-trained to perform masked-token prediction (BERT) or next-token prediction (GPT) on large corpora of generic text (Wikipedia, Common Crawl, etc.). Their complex architectures allow for rich interdependencies among tokens. As their size and complexity grow, so does these models' ability to perform sophisticated NLP tasks like question answering and document summarization. **GPT-3**, for example, is a massive neural network with 175B estimated parameters, and more recent models are even larger (e.g. **PALM** by Chowdhery et al. 2022).

While this size and complexity has resulted in stunning performance on NLP tasks, a downside is that these large models lack transparency and clear statistical structure. Training models with hundreds of billions of parameters requires vast hardware resources. Only large organizations can afford these, so most researchers must begin by downloading previously fitted models and potentially updating them. Hence, while it is possible to reuse the pre-trained models, replicating the full estimation pipeline is not possible.

2.6 Supervised learning for text

The algorithms discussed so far do not incorporate document metadata, but this is often of interest in economics applications. One instance is the supervised learning problem of predicting an outcome variable y_d (e.g. economic conditions or political affiliations) given $\mathbf{w}_d$. A straightforward approach to this problem is to use the bag-of-words model (potentially incorporating n-grams) and apply off-the-shelf high-dimensional regression models to estimate $\mathbb{E}[\, y_d \mid \mathbf{x}_d \,]$. The familiar penalized linear models in economics, such

[15] See Phuong and Hutter (2022) for a more extensive and formal description of Transformer models.

as LASSO, are typically too limited for text-related prediction tasks because they ignore the strong dependency structure in $\mathbf{x}_d$. Approaches such as random forests and gradient boosting are more robust in this environment since they allow for rich non-linearities and interactions among term counts (Hastie et al. 2009).

When deciding among supervised learning models, another relevant consideration is the corpus size D. Whereas computer science applications can have millions of labeled observations, social science applications might only have a few hundred. Ng and Jordan (2001) argue that joint models $p(y_d, \mathbf{x}_d)$ have worse asymptotic prediction error than conditional models $p(y_d \mid \mathbf{x}_d)$ but reach their asymptotic limit faster. Hence, datasets with relatively few observations might benefit from modeling this additional structure. Some example methods in this vein are supervised LDA (Mcauliffe and Blei 2007) and multinomial inverse regression (Taddy 2013, 2015).

A deeper issue is that using term counts for prediction rules out local interactions between terms. As we saw above, and as emphasized in modern NLP, a word's relevance often cannot be separated from the context of surrounding words. As such, there are corresponding benefits to adapting sequence-embedding methods for supervised learning. In the standard workflow, pre-trained models are "fine-tuned" for a supervised learning task—that is, a network trained for language-based prediction tasks is updated for a different prediction task. Appendix Figure B.1 in the supplementary materials illustrates the attention weights for a pre-trained BERT model that has been fine-tuned to predict whether a job posting offers remote work. Such an approach will usually dominate bag-of-word-based models and can approach human performance. Further, because the pre-trained Transformer models have a quite general understanding of diverse texts, fine-tuning can achieve good performance even with relatively few labeled training samples.

One limitation of BERT-based models is limited interpretability, which we discuss further in Section 5. Another is that they only operate on relatively short documents, typically 512 word-piece tokens, or about 400 words. This works well for sentences or paragraphs, but not for longer documents such as political speeches, judicial opinions, or corporate filings. While there are efforts to build Transformer-based models that take in longer inputs (e.g. Beltagy et al. 2020, Zaheer et al. 2020), it is often better with long documents to use non-Transformer-based alternatives such as gradient boosting applied to $\mathbf{x}_d$ (as mentioned above).[16]

[16] Another option for long documents is the model from Joulin et al. (2016), a neural network that produces n-gram embeddings and averages them across the document before being input to a standard feedforward neural net for classification or regression. A downside of this model is that it requires a large number of labeled documents to work well.

3 Four Measurement Problems

The adoption of text algorithms in economics is primarily motivated by applied researchers' need to solve specific measurement problems rather than an interest in the structure of the algorithms *per se*. Here we discuss four common measurement tasks and how the algorithms reviewed in the previous section can address them.

3.1 Problem I: Measuring document similarity

Computing document similarity is a core task in NLP, underlying search engine output, recommendation systems, and plagiarism detection. In economics, the distance between two documents can be used to proxy the distance in some economically relevant space. One leading example is the work of Hoberg and Phillips (2010, 2016), who use the overlap in firms' product descriptions in regulatory filings to measure the degree to which they are competitors.

All methods for computing document similarity begin with some vector representation of documents. The standard distance measure used to compare vectors in text analysis is *cosine similarity*. Formally, the cosine similarity between vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ is $\frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{|\mathbf{v}_1||\mathbf{v}_2|}$, that is, the Pearson correlation computed across demeaned vectors. It is higher when the angle between two vectors is smaller, i.e. when they share similar directions in the vector space. This metric ensures similarity is driven by similar word use rather than document length, as would be the case with Euclidean distance.

The question then becomes how to form document vectors, and the algorithms above provide many options. The simplest method uses the bag-of-words count vector $\mathbf{x}_d$ directly. Another popular option is to use **term frequency-inverse document frequency** (**tf-idf**) weighting in which the raw $x_{d,v}$ counts are multiplied by

$$\text{idf}_v = \log\left(\frac{D}{\sum_d \mathbb{1}(x_{d,v} > 0)}\right) \tag{6}$$

which upweights words that are specific to certain documents (e.g. Manning et al. 2008). Examples of bag-of-words-based approaches to similarity include Cagé et al. (2020), who use the distance between online news articles and social media posts to group items into common stories; Kelly et al. (2021b), who analyze the novelty and influence of technologies using pairwise comparisons between U.S. patent filings; and Biasi and Ma (2022), who measure similarity between college syllabi and academic journal articles to proxy the gap between course content and the newest research.

Because the vocabulary size V is typically very large, and the count vectors $\mathbf{x}_d$ are typically sparse, the distance between the vectors $\mathbf{x}_{d'}$ and $\mathbf{x}_{d''}$ can be a highly noisy measure of heterogeneity between documents d' and d''. In such environments, some

form of dimensionality reduction is often employed. For example, Iaria et al. (2018) use LSA to quantify the overlap between scientific research agendas as measured from article titles, while Bertrand et al. (2021) use it to compare the content of policy proposal comments in the U.S. federal rulemaking process. Hansen et al. (2018) use LDA applied to U.S. Federal Reserve committee transcripts to measure policymakers' herding behavior following an increase in transparency.

Another approach uses word embeddings to represent documents. In this case, the vector for document d is $\frac{1}{N_d}\sum_n \hat{\boldsymbol{\rho}}_{w_{d,n}}$, i.e. the average over the word embeddings corresponding to words in the document. Hansen et al. (2021) use this method to detect the presence of skills in job descriptions for executive managers by comparing them with O*NET task descriptions. Kogan et al. (2019) use a similar approach, but with a tf-idf-weighted instead of simple average, to measure the extent to which occupations are exposed to technology as proxied by similarity of O*NET task descriptions with patent text. Acemoglu et al. (2021) use **Doc2Vec**, an embedding variant by Le and Mikolov (2014) to represent documents rather than words, to measure the influence of the heads of Chinese research departments by comparing the associated scientific papers.

This variety of methods for creating document vectors raises the issue of which approach should be preferred. We return to this important question in Section 5 where we compare methods on the same similarity task.

3.2 Problem II: Concept detection

Textual data provides a rich—and sometimes the *only*—source of information about many economically crucial concepts. Examples include economic policy uncertainty (Baker et al. 2016), skill demand in the labor force (Deming and Kahn 2018), economic sentiment (Shapiro et al. 2020), and technology adoption (Bloom et al. 2021). An important measurement problem is thus how to detect the presence of a concept in economic text.

3.2.1 Pattern matching

A standard approach is to employ **dictionary methods** within the bag-of-words model. A researcher specifies a term set $\mathfrak{D}$ whose elements relate to the concept. Each document can be represented as the count over matched terms $z_d = \sum_{v\in\mathfrak{D}} x_{d,v}$, although many variants exist.[17] To specify these term sets, one has three common options. First, one can use sets derived from external sources. Enke (2020) applies a dictionary of moral val-

[17]For example, one can match on a binary indicator $z_d = \sum_{v\in\mathfrak{D}} \mathbb{1}(x_{d,v} > 0)$ or normalize by document length N_d. One can also use multiple dictionaries in combination to isolate a concept. Baker et al. (2016) use three terms sets to detect the presence of economic policy uncertainty (EPU) in individual newspaper articles: a set of economic terms, a set of uncertainty terms, and a set of policy-related terms. Newspapers articles are tagged as containing EPU language if they contain a term from each set.

ues terms built by social psychologists (the Moral Foundations Dictionary) to analyze a communal-vs-universalist dimension in congressional speeches. Hassan et al. (2019) build dictionaries of political language based on phrases' simultaneous presence in political science textbooks and absence in general financial language.[18] Second, one can use domain expertise to build term sets from scratch, such as the financial sentiment dictionaries of Loughran and Mcdonald (2011). Third, one can choose terms based on their ability to predict human-annotated documents (Baker et al. 2016, Advani et al. 2021).

Similar in spirit, but more general, than term-matching methods are pattern searches that use additional linguistic annotations besides words or characters. For one, a matching query could use a document's *part-of-speech* tags, to distinguish (for example) the noun "police" from the verb "police".[19] Further, *syntactic dependency* tags identify the connections between words – for example, which noun is the subject and which is the object.[20] Ash et al. (2020c) extract syntactic dependencies from labor union contracts to extract modal verbs (e.g., "shall", "may"), which work to impose obligations or specify permissions. Fetzer (2020) applies a syntax approach to detect and measure conflict events in a corpus of news articles from India.[21]

3.2.2 Algorithmic approaches

Moving beyond pattern matching, some of the algorithms discussed in Section 2 can be used to associate documents with concepts. Algorithms can help automate the construction of term sets, a task in which few economists have particular expertise even when they are clear on the concept they wish to measure. Algorithms can also help uncover more complex semantic rules for identifying concepts than those captured by term frequencies.

Topic model outputs. The dimensionality reduction algorithms in Section 2.3 automate the detection of latent concepts in a corpus and the words associated with these concepts. Take the literature on central bank communication, for example: Boukus and Rosenberg (2006) use LSA and Hansen and McMahon (2016) use LDA to decompose public documents released by central banks to study how specific topics relate to market movements. In forecasting, several recent papers have applied LDA to newspaper corpora

[18] Similarly, Mastrorocco and Ornaghi (2020) scan for municipality names in newspaper articles using a prepared list to identify mentions of places, and they also detect crime-related news stories by the presence of bigrams that are distinctive of those stories in a tagged corpus.

[19] Parts of speech tags identify the grammatical functions of words. A more sophisticated, but especially useful, tagging algorithm is *named entity recognition*, which works to identify references to specific people, organizations, or places (e.g. Jurafsky and Martin 2020).

[20] The relevant algorithm is called a syntactic dependency parser, which identifies head-dependent connections between words in a hierarchical tree structure.

[21] Moving beyond syntax to semantics, Ash et al. (2021b) use linguistic annotations for agents – the actors – and patients – the targets of actions – to construct and quantify micro-narratives. They then analyze the expressed connections between entities in partisan narratives used by U.S. legislators.

and interpreted the content of topics in terms of economic phenomena (Mueller and Rauh 2018, Larsen and Thorsrud 2019, Thorsrud 2020, Bybee et al. 2021).

An inherent challenge in unsupervised dimensionality reduction algorithms is that they do not generate objective topic labels. A given topic consists of many words, and words are scattered across many topics, so the outputs are often difficult to interpret. Even when the topic outputs are interpretable, unsupervised learning tools are wholly data-driven and cannot be targeted toward identifying specific concepts. This can be a strength in situations where the researcher wishes to explore the content of a corpus without strong prior knowledge. But when the goal is to link topics to specific economic concepts, the algorithm itself cannot achieve this. Objective interpretation is complicated by the fact that topic models can be sensitive to particular preprocessing and modeling choices (e.g. Denny and Spirling 2018b).

Given these concerns, one use of topic models is to provide an initial filter to remove clearly unrelated content, and then use more targeted methods to measure concepts in the remainder. Angelico et al. (2022) use this strategy to identify the relevant set of tweets for measuring inflation expectations.

Dictionaries augmented with machine judgment. Another approach is to specify an initial set of "seed" words that reflect a concept and then to use word embeddings to further populate the set with terms near the seeds in the associated vector space. Given an initial set of seed words, one can use cosine similarity between word vectors to either retrieve the nearest neighbors of the average embedding vector or to each seed individually. Researchers can then choose which of the retrieved words to include in the expanded term set.[22]

This approach allows the researcher to retain discretion on which concepts to measure while leveraging algorithms to map out how concepts relate to individual vocabulary terms. It is increasingly popular in a number of macro and finance applications (Hanley and Hoberg 2019, Atalay et al. 2020, Davis et al. 2020, Li et al. 2021, Soto 2021). In political economy, Gennaro and Ash (2021) use this method to populate 'emotionality' and 'rationality' term sets, and also to remove words that are outliers according to cosine similarity. Truffa and Wong (2022) use a word similarity algorithm to generate additional terms related to women and females, to help detect academic articles pertaining to gender.

These methods do not address the issue of polysemy—that is, words with multiple meanings. The word 'bank' for example could refer to a financial institution or a riverbank. Word embeddings – especially pre-trained embeddings learned from generic

[22] A similar approach can be used to improve the interpretability of topic models. CorEx (Gallagher et al. 2017) allows the researcher to nudge the topic model toward finding particular topics by providing seed words. Djourelova et al. (2021) use CorEx to help identify interpretable local news topics in their analysis of how CraigsList affected newspapers.

corpora like Wikipedia – will have vectors that combine both senses of the word. Given a specific mention of 'bank', a human can easily distinguish which of these meanings is more relevant. Recent embedding algorithms like ELMo (Peters et al. 2018), or a la carte embeddings (Khodak et al. 2018), draw on the neighboring words to produce context-sensitive embeddings that distinguish word senses. These embeddings solve the polysemy problem but increase design and computational complexity.

Embedding similarity of documents to word lists. Dictionaries generally provide coarse, lumpy measures of a concept of interest. They might not contain all semantically relevant terms, and terms are not graded by the intensity of their relationship with a concept. In the case of sentiment, for example, the word 'fantastic' will be coded the same as 'good' by a dictionary model. To address such issues, a researcher might want a more continuous scalar measure.

Again, word embeddings can address this issue. The idea is to put both the documents and the dictionary into the semantic space defined by the word embeddings, and then compute the proximity of each document to the dictionary. The simplest approach is to represent the dictionary and individual documents as average embedding vectors, and compute the cosine similarity between each document and the dictionary to obtain a continuous measure of association. Variants of this approach weight word vectors by inverse document frequency (e.g. Arora et al. 2016).

Ash et al. (2020a) apply this idea to measure the use of economics language by judges. They compute the similarity between embedded representations of the text of individual judges and a lexicon of economics-related phrases. Judges who attend economics training use more economics language. Gennaro and Ash (2021) produce embedding dimensions for emotion and rationality dictionaries, and then scale political speeches along an emotionality index by their relative distance to these dictionary vectors. They then explore the use of emotional rhetoric in speeches by U.S. Congress Members.

Machine prediction based on human annotations. The most accurate approach to concept detection is perhaps direct human reading with appropriate domain expertise. But labeling all documents can be too costly in time and money. Hence, a common strategy is to use human reading on a subset of data to generate labels and to treat concept detection as a supervised learning problem targeting those labels (Adams-Prassl et al. 2020, Besley et al. 2020, Shapiro et al. 2020). The prediction model is then taken out-of-sample to impute human labels, effectively scaling up human reading to the whole corpus. The main consideration is then building an accurate regression model, where the methods described in Section 2.6 can be directly implemented. Even when the goal is not to use supervised learning methods directly, human labels can be useful to benchmark the

performance of dictionaries and to guide the choice of which particular terms to include, as in Baker et al. (2016).

Hansen et al. (2022) compares several supervised learning models for predicting human labels for remote work, and find that BERT-like models achieve outstanding performance. The intuition is that sequence embedding models can use the context around terms to determine whether they flag the relevant concept. For example, both of the following sentences would be flagged as offering remote work under a naïve dictionary search for the term 'remote work':

1. "This position involves travel to remote work sites."

2. "Remote work is supported under our work-from-home policy."

but only 2 is a correct flag. Separating out these cases requires going beyond word counts, word associations, or syntactic patterns and instead modeling how words in language interrelate to generate meaning. Attention-based classifiers excel at these complex tasks. Of course, with greater predictive power comes a decrease in interpretability and how to resolve this trade-off will vary between applications.

3.3 Problem III: How concepts are related

The third problem we consider is how concepts are related in a corpus: for example, positive or negative sentiment with economic conditions (Apel and Blix Grimaldi 2014); risk with political exposure (Hassan et al. 2019); and career and family with gender (Ash et al. 2020b). The simplest approach begins from dictionaries that represent two concepts of interest, then tabulates the number of times terms from each dictionary co-occur within a local window (Apel and Blix Grimaldi 2014, Hassan et al. 2019, Cieslak et al. 2021).

Many variants of this basic approach exist and build on algorithms from Section 2. When one has a strong prior on one concept of interest but a weak one on the other, dictionaries and topic models can be combined. For example, Larsen and Thorsrud (2019) and Thorsrud (2020) estimate LDA on a large Norwegian financial newspaper and group articles most associated with each topic. Then they apply sentiment dictionaries to each separate topical group.[23]

As mentioned above, dictionary counts produce coarse representations of concepts. The local co-occurrence method exacerbates this problem because it requires simultaneous mentions of terms from two sets which can lead to sparse measures. The *word embedding association test* (WEAT, Caliskan et al. 2017) addresses this problem with word embeddings. It begins with sets of attribute words A and B that denote opposite

[23] Vafa et al. (2020) present a full generative model that captures the idea that a latent dimension (e.g. sentiment) interacts with the language inside topics.

ends of a conceptual spectrum. For example, A (B) might contain words reflecting positive (negative) sentiment. Then any other word, or set of words, can be projected into the conceptual space by measuring its relative position between A and B with cosine similarity. Figure 3 (from Kozlowski et al. 2019) locates various terms in two separate conceptual dimensions built with a U.S.-specific corpus. The term locations in the social class and left-right political dimensions are reasonable.

The first application of word-embedding associations in economics is Ash et al. (2020b), which measures gender attitudes of individual U.S. appellate court judges by applying WEAT separately to each judge's authored opinions and considering the correlation between male-female and career-family dimensions. Gender attitudes of judges relate to decisions and treatment of female colleagues.[24]

3.4 Problem IV: Associating text with metadata

In some situations, text comes with metadata that forms the basis of measurement. This strategy is particularly useful when one has a set of documents with an outcome variable of interest whose value needs to be imputed to other documents. One well-known example is Gentzkow and Shapiro (2010), who use the political party associated with speakers in the U.S. Congressional Record to build a regression model that maps speech into a predicted party label. They then use this model to attribute a political bias to media outlets based on the text of their articles, a form of supervised transfer learning. Similarly, Widmer et al. (2020) produce a measure of slant based on similarity of newspaper language to that used by Fox News Channel.

The appropriate tool for tackling this problem is supervised learning, as the goal is to maximize the goodness-of-fit in new documents. Hence, the considerations laid out in Section 2.6 can help guide the choice of model. When applying this approach, care must be taken that the unlabeled documents have the same association between words and outcomes as the training corpus. Osnabrügge et al. (2021) evaluate the performance of supervised transfer learning by assessing the extent to which out-of-sample predictions line up with human labels.

In other cases, supervised learning may be an end in itself without being used for outcome imputation. For example, Bana (2022) fine-tunes a BERT model to predict salaries from the text of job postings and performs counterfactual exercises on salary outcomes by varying the language input. Ke et al. (2019) and Davis et al. (2020) use text of news articles and regulatory filings, respectively, to predict stock returns using supervised learning models, which notably outperform standard dictionaries.

[24] Jha et al. (2022) defines attribute sets A and B with sentences instead of words to measure sentiment towards finance. They use BERT to project historical book extracts into finance sentiment space.

4 Text Measures and Econometric Models

Each of the four measurement strategies outlined above convert text into a quantitative measure. These papers do not stop after preparing the derived measures, but also use them as inputs in downstream econometric models. For example, Baker et al. (2016) include their dictionary-based EPU index in a VAR along with traditional macro data. In the context of monetary policy deliberations, Hansen et al. (2018) take LDA shares as features and analyze which ones respond to changes in central bank transparency. Widmer et al. (2020) use Fox News Channel position as an instrument for the popularity of the network, and show that in places with higher Fox News viewership, the local newspaper uses language that is more similar to Fox News than other cable news networks.

For the most part, text quantification algorithms and econometric models are treated separately, where the former create "data" that is treated like any other numeric covariate in the latter. This approach creates potential inference problems that the economics literature has hardly begun to explore but which are important to highlight.[25] For one, the downstream econometric model ignores uncertainty present in the upstream measurement. Also, shared dependencies are ignored which may add to the measurement error. For example, LDA assumes that all document-topic vectors are drawn independently and identically from a Dirichlet prior. Treating those vectors as depending on document-level metadata in follow-on regression models violates that assumption.

One solution is to allow the upstream measurement model and the downstream econometric model to communicate with each other within a single, integrated model. Statistical models of text provide a means of specifying the structure and conducting valid inference. Taddy (2013) and Taddy (2015) model the multinomial probability $\mathbf{q}_d$ in (1) with a (penalized) multinomial logistic regression that depends on document-level covariates. Gentzkow et al. (2019b) use this framework to connect word frequencies in Congressional speeches to political party affiliation and analyze historical variation in partisanship. Kelly et al. (2021a) extend this regression framework to account for the excess zeros present in the term counts for the bag-of-words model.

Meanwhile, LDA has been extended in many directions to jointly model latent topical structure and covariates. A leading example is the structural topic model (Roberts et al. 2014), which adjusts the prior distribution over $\boldsymbol{\theta}_d$ to account for covariate dependencies. One barrier to the adoption of these models is the complex Bayesian inference algorithms needed for posterior approximation, although recent breakthroughs in automatic inference relax these considerably (Sacher et al. 2021).

Still, many of the measurement approaches discussed do not have a well-defined statistical model for likelihood-based inference. There are examples of neural network models

[25] We refer readers to Grimmer et al. (2022) who provide an overview of some of the salient issues.

that have been adapted to incorporate covariate dependencies (Pryzant et al. 2018), but how one conducts valid inference with them is not clear. This is an area of active research (e.g. Farrell et al. 2021).

These more sophisticated structural and inference methods do not address more fundamental issues of identification, and more specifically of non-classical measurement error with text data. Text algorithms are applied with the goal of measuring some economically relevant dimension in text, but they might often bring in other correlated factors. As an example, a classifier trained to predict whether job posts involve remote work might learn that software development tends to be remote. Using such predictive information is not a problem for a static prediction task. But let's say we would like to estimate the treatment effect of a recession on remote work. We might estimate a spurious treatment effect that is due to how the recession affects the share of software development jobs, rather than its effect on remote work. The problem is an exclusion restriction violation – that estimated treatment effects could be biased by effects of the treatment on the confounding predictors, rather than the latent dimension of interest.

There are no simple solutions to this problem. If anything, the more sophisticated supervised learning algorithms, like BERT, are more vulnerable to it because they use more subtle style features in making predictions, such as punctuation. Dictionary methods are less likely to bring in correlated factors, but they have the other downsides discussed above. One essential validation check is to use an annotated sample to show that the model's error rate is uncorrelated with the treatment. If the treatment affects the model error, then an exclusion restriction violation is likely.

5 Issues and Challenges

This section follows up on two themes that have come up repeatedly: validation (5.1) and interpretability (5.2) in the use of text algorithms. The section concludes by assessing the prospects of large pre-trained language models (5.3).

5.1 Validation

A theme from the preceding sections is that different researchers have employed a variety of algorithms for tackling the core empirical applications involving text. The logic guiding these choices is often not clear, nor is the sequence of implicit and explicit calculations leading from a corpus of documents to a set of regression coefficients. This would not be a problem if there were consensus tools that always worked as expected. But there is no such consensus. The methods for text-as-data are too new and varied, with specific applications requiring specific adjustments. In the text-as-data world, we are quite far

from the shared expectations about data wrangling, summary statistics, identification checks, regression models, and specification checks that have come to characterize more mature subfields like applied microeconometrics (e.g. Angrist and Pischke 2009).

To illustrate how specific (and often unexamined) modeling choices can matter for outcomes, we return to the document similarity task from Section 3.1 and compare alternative methods for comparing the similarity between documents from a popular corpus: *Risk Factors* language from annual 10-K filings. We use a sample of 4,033 firms for which we can obtain these texts based on 2019 filings. Pre-processing of the documents and further implementation details are described in the supplementary materials. To compare similarity, we use eleven different approaches to construct document vectors, all of which have appeared in the literature:

- Bag-of-words-based term counts: 1) raw counts; 2) tf-idf weighted term counts.
- Average word embedding based on: 3) pre-trained GloVe (pre-trained on Wikipedia) 4) GloVe estimated on the Risk Factors corpus; 5) same as 4 but using tf-idf weights to compute average; 6) word2vec estimated on the Risk Factors corpus; 7) same as 6 but using tf-idf weights to compute average.
- 8) Doc2Vec.
- Dimensionality reduction of document-term matrix: 9) LSA, 10) NMF, 11) LDA.

We first compute pairwise cosine similarities across firms according to each method. Figure 4a shows the Pearson correlation between the obtained similarities. To compare ordinal rankings, we also draw 10,000 document triplets and use each approach to ask whether the second or third document is closest to the first. Figure 4b shows the fraction of cases in which the methods agree. While some of the embedding-based approaches show high agreement with each other, in general there is large divergence across methods. The average Pearson correlation across the reported cells in Fig. 4a is just 0.57, while the average agreement rate from Fig. 4b is just 0.74.

In the supplementary materials, we describe a similar exercise for word comparisons using four different word embeddings models. Appendix Figure B.3 is the analogue of Figure 4 and shows at least as large divergences across the four algorithms. For word similarities, the average Pearson correlation across algorithms is just 0.42, with a 0.64 agreement rate when ranking triplets.

These divergences would not be problematic if they arose from random noise uncorrelated with economic fundamentals. To assess this, we draw a random sample of 50,000 firms pairs and again compute cosine similarities using each method. We then regress each set of pairwise similarities on a set of covariates comparing the firms: whether the firms share a NAICS2 sector; the correlation between 2019 daily stock returns; and the

difference in firm size as measured by the absolute log ratio of employees and, additionally, of total assets.

Figure 5 displays the estimated effects, where the dependent and (continuous) independent variables are in standard deviation units. While most estimated effects go in the expected direction, point estimates and confidence intervals differ greatly, and methods disagree on which covariate is most associated with textual similarity. Hence, in this application—which is emblematic of many in the literature—the choice of algorithm is not innocuous for downstream inference. Given the battery of specification and robustness checks that accompany applied research, it is notable how little attention upstream modeling choices receive.

How should one proceed? Establishing which algorithm best captures the overlap in economically relevant risk factors is impossible without further information.[26] The information retrieval and NLP literatures have established standard external evaluation tasks to judge the performance of algorithms. For document similarity, for example, one could use search engine click-through rates as a measure of the relevance of a document ranking for users. For word similarity, the NLP literature would typically use tasks like synonym detection or analogy completion. But in economics we currently lack such objective benchmarks against which to validate the choice of algorithm.

A major step forward in the text-as-data literature would be to discipline modeling choices by assessing their performance on standardized tasks our field views as important. To the extent that researchers validate algorithms at all, they tend to present a few promising examples of algorithmic output *ex post*, which is potentially prone to researcher manipulation. A limited number of papers perform more rigorous *ex post* assessments. Gennaro and Ash (2021) perform an extensive validation exercise to determine whether word-embedding-based measures of concepts correspond with human judgment. Lippmann (2022) inspects all of the legislative amendments tagged by his dictionary as gender-related, to check high precision. While these are valuable exercises, they are specific to the respective settings and provide little guidance for setting more general *ex ante* criteria against which multiple algorithms could be compared.

Some field-specific text-data validations are easy to imagine. For word embeddings, one could define economic-specific word relationships that would be desirable for a model to resolve and then ask which model comes closest to doing so.[27] For example, economically relevant word embeddings would be able to complete the analogy 'CPI is to inflation as GDP is to [MASK]' by filling [MASK] with 'output'.[28] Such basic annotations based

[26] There may be generic statistical arguments for why one approach is preferred, but such explanations tend to be *ad hoc* and unrelated to the economic environment.

[27] See Rodriguez and Spirling (2020) who conducts this exercise in a political science context.

[28] Generic pre-trained embedding models might do a poor job at representing economic ideas. The GloVe vectors estimated on Wikipedia produce as nearest neighbors to 'team' words like 'squad', 'players',

on economic reasoning could be done even by undergraduate economics majors.

Validations requiring deeper expert judgment on economic matters are also worth exploring. In the corporate filing context, for example, experts could be asked to code the specific risks present in a subset of filings, where the codebook could be informed by economic and financial models. This annotation would produce data on which firms shared similar risks, and one could ask which of the eleven models mapped these firms into similar vectors. Such expert assessments may be too subjective to be reproducible, however, or they may be too costly.

In any case, some human input is needed. Given the time and expense of developing annotations for validation purposes, the incentives for individual research teams are to develop smaller-scale validations that fit a specific project. But allowing each paper to design its own validation task brings one back to the problem of having no common standard. A more effective long-term approach is to produce validation tasks that are specific to economics but relevant to a broad range of economics applications. Ideally, a battery of *standardized* validation tasks could be developed by the text-as-data community to provide the profession with an objective benchmark for modeling choices. Much of the success of the NLP literature in computer science has been in the development of generic tasks for language models, such as GLUE (Wang et al. 2018), which provide a set of benchmarks for diverse language tasks and help motivate measurable progress. The creation of a similar resource for the text-as-data community in economics might enable analogous breakthroughs. Whether and how such ex-ante validation baselines could be developed, and whether they would actually work in practice, remains to be seen.

A number of other recent technical developments can facilitate the development of these validation baselines. First, the human annotation process can be sped up by machine support, for example *active learning* procedures where documents are sorted for labeling by their usefulness in reducing the entropy of the classifier's predicted probabilities (Monarch 2021). Another promising set of approaches in the area of *weak supervision*, where labels generated automatically by the environment can be combined with minimal human supervision to label large document collections (Sedova et al. 2021). Finally, as discussed further below, large pre-trained language models like GPT-3 can help by machine-labeling documents.

5.2 Interpretability

A repeated theme of our discussion is a trade-off between performance (i.e. label prediction accuracy) and interpretability. Typically, the best approach in terms of performance is to use a sophisticated Transformer-based classifier that best predicts the variable in

'football', and 'coach'. In economics, the word 'team' is used more often in discussions of production and firm organization than in discussions of sports.

a held-out sample. However, this choice may not be the most interpretable: attention-based, deep neural network excel at supervised learning but are notoriously opaque.

There are two reasons economists might care about interpretability. First, if a predictive model is deployed to understand mechanisms, good prediction is not enough. Consider the problem of predicting speakers' political ideology from their speeches. 'Texas' might be an accurate and influential predictor of right-wing ideology but is not a term structurally related to a belief system. More relevant terms for defining right-wing ideology (in the 2022 U.S. context) would relate to small government, the importance of religion, immigration restrictions, etc. The issue is that 'Texas' is likely to co-occur with such terms and be used more often by right-wing speakers. Due to the high dimensionality inherent in its feature space, text is prone to generate many such spurious correlations which predictive models will nonetheless use to achieve good fit.

In terms of understanding mechanisms, interpretability is also a central criterion in model selection for unsupervised learning. With LDA, for example, one has to decide the number of topics K. Chang et al. (2009) choose K based on human judgments about topic coherence – specifically, the rate at which annotators correctly identify an 'intruder' term that has been randomly inserted into each topic's list of most-associated terms. Similarly, Demszky et al. (2019) set the options for a tweet clustering algorithm using an intruder detection task. In general, the topic number that maximizes humans' ability to interpret the output of unsupervised-learning models diverges from the number that maximizes goodness-of-fit in held-out data.

The second reason is that predictive performance on existing data may not be representative of performance in new domains. For example, an algorithm for predicting recessions from newspaper articles through 2020 might miss the 2022 downturn due to the novel features of the latter. Economic data is subject to considerably more noise and structural breaks than data typical of the environments in which modern NLP algorithms were originally developed. A reasonable hypothesis is that more complex models may prove less effective for prediction when outcomes are drawn from new distributions.

One solution to these problems is to use simple approaches, such as dictionary methods or logistic regression with a small vocabulary (Rudin 2019), where one can relatively easily understand the algorithm's classification logic. These simple models will generally perform worse at predicting labels, however (Kleinberg and Mullainathan 2019). Second, one can use model explanation methods to provide interpretable diagnostics on the features that an algorithm is relying on (Ribeiro et al. 2016). These methods can diagnose cases where models are relying disproportionately on spurious correlates, but do not immediately deliver a solution. The simplest response is to then pre-process documents to remove those correlated features, but the consequences of such targeted pre-processing have not been systematically examined.

As the economics literature using text progresses, new methods and approaches to resolving the tension between prediction and interpretability will be needed. Because NLP has moved in the direction of developing ever-more-complex models, this issue is arguably even more salient than for other machine learning approaches in economics.

5.3 Possibilities of large pre-trained language models

On a more forward-looking note, it is worth revisiting the intriguing and powerful properties of pre-trained language models such as BERT and GPT-3. One immediately useful application is to incorporate multilingual text into empirical analysis. To date, most text analysis in economics has been in English only, a situation pre-trained language models can help overcome. High-performing machine translation systems are now available as open-source packages (e.g. Tiedemann and Thottingal 2020). Further, recent generations of document encoders are built multilingually, such that semantically equivalent documents in different languages are given the same vector representation (e.g. Artetxe and Schwenk 2019).

More speculatively, to the extent that models like GPT-3 understand language, they may be able to take over language-related research tasks in economics. As an example, consider this paper's abstract. Its text was written purely by GPT-3, with the introduction as input accompanied by a prompt to "Write an abstract for the following scientific article".[29] The authors have used GPT-3 in other similar tasks, such as generating paper titles. In the accompanying notebook, we illustrate this power of GPT-3 to generate fluent academic writing.

Beyond support for writing papers, there is also proven performance for language models in writing code (e.g. Xu et al. 2022). It is not improbable that these language models will work for supporting code development for economics research, including data wrangling and regression analysis. These models will be able to generate well-formatted tables and other result reporting modalities. More uncertain is to what extent such language models will be able to read and evaluate research outputs, for example to support the peer review process.

More specific to text as data, GPT-3 and related models will be useful in data labeling and validation. For well-defined language tasks, GPT-3 achieves human performance. So for example, labeling documents as belonging in a category, or having some feature, should be possible with GPT-3, perhaps with additional human supervision. Stammbach et al. (2022) illustrate this possibility in the case of identifying narrative roles in texts—that is, extracting heroes, villains, and victims from plain-text stories.

[29] We used the 'text-davinci-002' model from the OpenAI API, with temperature = 0.7, frequency penalty = 0, and presence penalty = 0. LaTex code was removed from the input text.

6 Conclusion

To summarize, text algorithms provide a window into many interesting research questions for economists, although they do not immediately solve the challenges of text data. There are effective tools for transforming strings to vectors, for informatively reducing the dimensionality of those vectors, and for leveraging that information for regression and other tasks. The system outputs can be used for measuring and comparing documents, the economic variables represented in the documents, and the metadata associated with the documents.

Future work could expand text analysis in economics in a number of directions. First, as discussed above, coordinating on a well-defined set of validation tasks would give economists a more principled foundation for choosing among the many available models. Second, building new models that jointly model text and numeric data could help address some of the inference problems that arise from plugging algorithmic output from off-the-shelf NLP into downstream regression models. Third, exploring the uses of text in causal inference is an important next step once the core measurement problems are addressed.[30] Fourth, text has almost exclusively been used in reduced-form empirical exercises, but in principle can also inform structural estimation. Finally, large pre-trained language models may be relevant in many research tasks, including labeling data or even to help write research papers.

The algorithms we discuss in this review, or close variants, are also useful for representing other unstructured datasets beyond text. Bandiera et al. (2020) use LDA to measure leadership styles of CEOs from a detailed time use survey (see also Draca and Schwarz 2018). Ruiz et al. (2020) use a model related to word embeddings to capture latent characteristics of goods that generate co-occurrence patterns in customer shopping baskets. Ash et al. (2021a) use images of individuals in newspapers to map out patterns of visual bias. These initial explorations point toward a broader base of unstructured data for economists to draw on in the coming decades.

[30] One example in this direction is Ash et al. (2020d), who construct a shift-share instrument for the volume of legislative output in U.S. states using a topic model. Analogous with shift-share instruments for economic output that multiply pre-period local sectoral shares with current-period leave-one-out national sectoral shocks, the legislative instrument is constructed as the pre-period local topic shares in state legislation times the current-period leave-one-out national shocks by topic. Using the instrument, that paper shows that higher legislative output caused higher economic output in recent decades.

References

Acemoglu, D., Yang, D. Y., and Zhou, J. (2021). Political pressure and the direction of research: Evidence from chinas academia. Technical report, Working paper.

Adams-Prassl, A., Balgova, M., and Qian, M. (2020). Flexible Work Arrangements in Low Wage Jobs: Evidence from Job Vacancy Data. *SSRN Electronic Journal*.

Advani, A., Ash, E., Cai, D., and Rasul, I. (2021). Race-related research in economics and other social sciences. *Econometic Society Monograph Series*.

Angelico, C., Marcucci, J., Miccoli, M., and Quarta, F. (2022). Can we measure inflation expectations using Twitter? *Journal of Econometrics*, 228(2):259–277.

Angrist, J. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An empiricist's companion*. Princeton University Press, Princeton, NJ.

Apel, M. and Blix Grimaldi, M. (2014). How Informative Are Central Bank Minutes? *Review of Economics*, 65(1):53–76.

Arora, S., Liang, Y., and Ma, T. (2016). A simple but tough-to-beat baseline for sentence embeddings.

Artetxe, M. and Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Ash, E., Chen, D., and Naidu, S. (2020a). Ideas have consequences: The impact of law and economics on american justice. *Center for Law & Economics Working Paper Series*, 4.

Ash, E., Chen, D. L., and Ornaghi, A. (2020b). Gender attitudes in the judiciary : Evidence from U.S. circuit courts. https://warwick.ac.uk/fac/soc/economics/research/workingpapers/2020/twerp_1256_-_ornaghi.pdf.

Ash, E., Durante, R., Grebenschikova, M., and Schwarz, C. (2021a). Visual stereotypes in news media.

Ash, E., Gauthier, G., and Widmer, P. (2021b). Text semantics capture political and economic narratives. *arXiv preprint arXiv:2108.01720*.

Ash, E., Jacobs, J., MacLeod, B., Naidu, S., and Stammbach, D. (2020c). Unsupervised extraction of workplace rights and duties from collective bargaining agreements. In

2nd International Workshop on Mining and Learning in the Legal Domain (MLLD-2020)(virtual).

Ash, E., Morelli, M., and Vannoni, M. (2020d). More laws, more growth? evidence from us states. *Center for Law & Economics Working Paper Series*, 15.

Atalay, E., Phongthiengtham, P., Sotelo, S., and Tannenbaum, D. (2020). The Evolution of Work in the United States. *American Economic Journal: Applied Economics*, 12(2):1–34.

Bahdanau, D., Cho, K. H., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636.

Bana, S. H. (2022). work2vec: Using language models to understand wage premia.

Bandiera, O., Prat, A., Hansen, S., and Sadun, R. (2020). CEO Behavior and Firm Performance. *Journal of Political Economy*, 128(4):1325–1369.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Bertrand, M., Bombardini, M., Fisman, R., Hackinen, B., and Trebbi, F. (2021). Hall of mirrors: Corporate philanthropy and strategic advocacy. *The Quarterly Journal of Economics*, 136(4):2413–2465.

Besley, T., Fetzer, T., and Mueller, H. (2020). How big is the media multiplier? evidence from dyadic news data. *Unpublished manuscript*.

Biasi, B. and Ma, S. (2022). The education-innovation gap. Technical report, National Bureau of Economic Research.

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

Bloom, N., Hassan, T. A., Kalyani, A., Lerner, J., and Tahoun, A. (2021). The Diffusion of Disruptive Technologies. Working Paper 28999, National Bureau of Economic Research.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Boukus, E. and Rosenberg, J. V. (2006). The Information Content of FOMC Minutes.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Bybee, L., Kelly, B. T., Manela, A., and Xiu, D. (2021). Business News and Business Cycles.

Cagé, J., Hervé, N., and Viaud, M.-L. (2020). The Production of Information in an Online World. *The Review of Economic Studies*, 87(5):2126–2164.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., and Blei, D. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2022). PaLM: Scaling Language Modeling with Pathways.

Cieslak, A., Hansen, S., McMahon, M., and Xiao, S. (2021). Policymakers' Uncertainty.

Coase, R. H. (1960). The Problem of Social Cost. *The Journal of Law & Economics*, 3:1–44.

Davis, S. J., Hansen, S., and Seminario-Amez, C. (2020). Firm-Level Risk Exposures and Stock Returns in the Wake of COVID-19. Working Paper 27867, National Bureau of Economic Research.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Deming, D. and Kahn, L. B. (2018). Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals. *Journal of Labor Economics*, 36(S1):S337–S369.

Demszky, D., Garg, N., Voigt, R., Zou, J., Gentzkow, M., Shapiro, J., and Jurafsky, D. (2019). Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. *arXiv preprint arXiv:1904.01596*.

Denny, M. J. and Spirling, A. (2018a). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis*, 26(2):168–189.

Denny, M. J. and Spirling, A. (2018b). Text Preprocessing for Unsupervised Learning: Why it Matters, when it misleads, and what to Do about it. *Political Analysis*, https://doi.org/10.1017/pan.2017.44.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ding, C., Li, T., and Peng, W. (2006). Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence, chi-square statistic, and a hybrid method. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pages 342–347, Boston, Massachusetts. AAAI Press.

Djourelova, M., Durante, R., and Martin, G. (2021). The impact of online competition on local newspapers: Evidence from the introduction of craigslist.

Draca, M. and Schwarz, C. (2018). How polarized are citizens? measuring ideology from the ground-up.

Enke, B. (2020). Moral values and voting. *Journal of Political Economy*, 128(10):3679–3729.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Fetzer, T. (2020). Can workfare programs moderate conflict? evidence from india. *Journal of the European Economic Association*, 18(6):3337–3375.

Friedman, M. and Schwartz, A. J. (1963). *A Monetary History of the United States: 1867-1960*. National Bureau of Economic Research. Princeton University Press, Princeton, NJ.

Gallagher, R. J., Reing, K., Kale, D., and Ver Steeg, G. (2017). Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5:529–542.

Gennaro, G. and Ash, E. (2021). Emotion and reason in political language. *Economic Journal*, 2021(02).

Gentzkow, M., Kelly, B., and Taddy, M. (2019a). Text as Data. *Journal of Economic Literature*, 57(3):535–574.

Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71.

Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019b). Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica*, 87(4):1307–1340.

Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

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

Grimmer, J., Roberts, M. E., and Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Grimmer, J. and Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297.

Hanley, K. W. and Hoberg, G. (2019). Dynamic Interpretation of Emerging Risks in the Financial Sector. *The Review of Financial Studies*, 32(12):4543–4603.

Hansen, S., Lambert, P., Bloom, N., Davis, S., Sadun, R., and Taska, B. (2022). Remote Work across Jobs, Companies, and Countries. Working Paper.

Hansen, S. and McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99:S114–S133.

Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach. *The Quarterly Journal of Economics*, 133(2):801–870.

Hansen, S., Ramdas, T., Sadun, R., and Fuller, J. (2021). The Demand for Executive Skills.

Hassan, T. A., Hollander, S., van Lent, L., and Tahoun, A. (2019). Firm-Level Political Risk: Measurement and Effects. *The Quarterly Journal of Economics*, 134(4):2135–2202.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Larning*. Springer, New York, NY.

Hoberg, G. and Phillips, G. (2010). Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis. *The Review of Financial Studies*, 23(10):3773–3811.

Hoberg, G. and Phillips, G. (2016). Text-Based Network Industries and Endogenous Product Differentiation. *Journal of Political Economy*, 124(5):1423–1465.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, pages 289–296, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Iaria, A., Schwarz, C., and Waldinger, F. (2018). Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science*. *The Quarterly Journal of Economics*, 133(2):927–991.

Jha, M., Liu, H., and Manela, A. (2022). Does Finance Benefit Society? A Language Embedding Approach.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Jurafsky, D. and Martin, J. H. (2020). *Speech and Language Processing. Third Edition*.

Ke, S., Olea, J. L. M., and Nesbit, J. (2021). Robust Machine Learning Algorithms for Text Analysis. Unpublished manuscript.

Ke, Z. T., Kelly, B., and Xiu, D. (2019). Predicting Returns with Text Data. page 55.

Kelly, B., Manela, A., and Moreira, A. (2021a). Text Selection. *Journal of Business & Economic Statistics*, 39(4):859–879.

Kelly, B., Papanikolaou, D., Seru, A., and Taddy, M. (2021b). Measuring Technological Innovation over the Long Run. *American Economic Review: Insights*, 3(3):303–320.

Khodak, M., Saunshi, N., Liang, Y., Ma, T., Stewart, B., and Arora, S. (2018). A la carte embedding: Cheap but effective induction of semantic feature vectors. *arXiv preprint arXiv:1805.05388*.

Kleinberg, J. and Mullainathan, S. (2019). Simplicity creates inequity: implications for fairness, stereotypes, and interpretability. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 807–808.

Kogan, L., Papanikolaou, D., Schmidt, L., and Seegmiller, B. (2019). Technology, Vintage-Specific Human Capital, and Labor Displacement: Evidence from Linking Patents with Occupations.

Kozlowski, A. C., Taddy, M., and Evans, J. A. (2019). The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*, 84(5):905–949.

Larsen, V. H. and Thorsrud, L. A. (2019). The value of news for economic developments. *Journal of Econometrics*, 210(1):203–218.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*.

Levy, O. and Goldberg, Y. (2014). Dependencybased word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308.

Li, K., Mai, F., Shen, R., and Yan, X. (2021). Measuring Corporate Culture Using Machine Learning. *The Review of Financial Studies*, 34(7):3265–3315.

Lippmann, Q. (2022). Gender and lawmaking in times of quotas. *Journal of Public Economics*, 207:104610.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loughran, T. and Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, illustrated edition edition.

Mastrorocco, N. and Ornaghi, A. (2020). *Who watches the watchmen? Local news and police behavior in the United States*. University of Warwick, Department of Economics.

Mcauliffe, J. and Blei, D. (2007). Supervised Topic Models. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546 [cs, stat]*.

Monarch, R. M. (2021). *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Mueller, H. and Rauh, C. (2018). Reading Between the Lines: Prediction of Political Violence Using Newspaper Text. *American Political Science Review*, 112(2):358–375.

Ng, A. Y. and Jordan, M. I. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, pages 841–848, Cambridge, MA, USA. MIT Press.

Osnabrügge, M., Ash, E., and Morelli, M. (2021). Cross-domain topic classification for political texts. *Political Analysis*, 4.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations.

Phuong, M. and Hutter, M. (2022). Formal Algorithms for Transformers.

Pryzant, R., Shen, K., Jurafsky, D., and Wagner, S. (2018). Deconfounded Lexicon Induction for Interpretable Social Science. In *Proceedings of the 2018 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1615–1625, New Orleans, Louisiana. Association for Computational Linguistics.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. pages 1135–1144.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4):1064–1082.

Rodriguez, P. L. and Spirling, A. (2020). Word embeddings.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Ruiz, F. J. R., Athey, S., and Blei, D. M. (2020). SHOPPER: A probabilistic model of consumer choice with substitutes and complements. *The Annals of Applied Statistics*, 14(1):1–27.

Rydning, J. (2021). Worldwide Global DataSphere and Global StorageSphere Structured and Unstructured Data Forecast, 2021–2025. Market Forecast, International Data Corporation.

Sacher, S., Battaglia, L., and Hansen, S. (2021). Hamiltonian Monte Carlo for Regression with High-Dimensional Categorical Data. *arXiv:2107.08112 [econ, stat]*.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*.

Sedova, A., Stephan, A., Speranskaya, M., and Roth, B. (2021). Knodle: Modular weakly supervised learning with pytorch. *arXiv preprint arXiv:2104.11557*.

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2020). Measuring news sentiment. *Journal of Econometrics*.

Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., and Li, W. (2021). Layoutparser: A unified toolkit for deep learning based document image analysis. In *International Conference on Document Analysis and Recognition*, pages 131–146. Springer.

Soto, P. E. (2021). Breaking the Word Bank: Measurement and Effects of Bank Level Uncertainty. *Journal of Financial Services Research*, 59(1):1–45.

Stammbach, D., Antoniak, M., and Ash, E. (2022). Heroes, villains, and victims, and gpt-3–automated extraction of character roles without training data. *arXiv preprint arXiv:2205.07557*.

Taddy, M. (2013). Multinomial Inverse Regression for Text Analysis. *Journal of the American Statistical Association*, 108(503):755–770.

Taddy, M. (2015). Distributed Multinomial Regression. *The Annals of Applied Statistics*, 9(3):1394–1414.

Thorsrud, L. A. (2020). Words are the New Numbers: A Newsy Coincident Index of the Business Cycle. *Journal of Business & Economic Statistics*, 38(2):393–409.

Tiedemann, J. and Thottingal, S. (2020). Opus-mt–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):611–622.

Truffa, F. and Wong, A. (2022). Undergraduate gender diversity and direction of scientific research.

Vafa, K., Naidu, S., and Blei, D. (2020). Text-Based Ideal Points. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5345–5357, Online. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

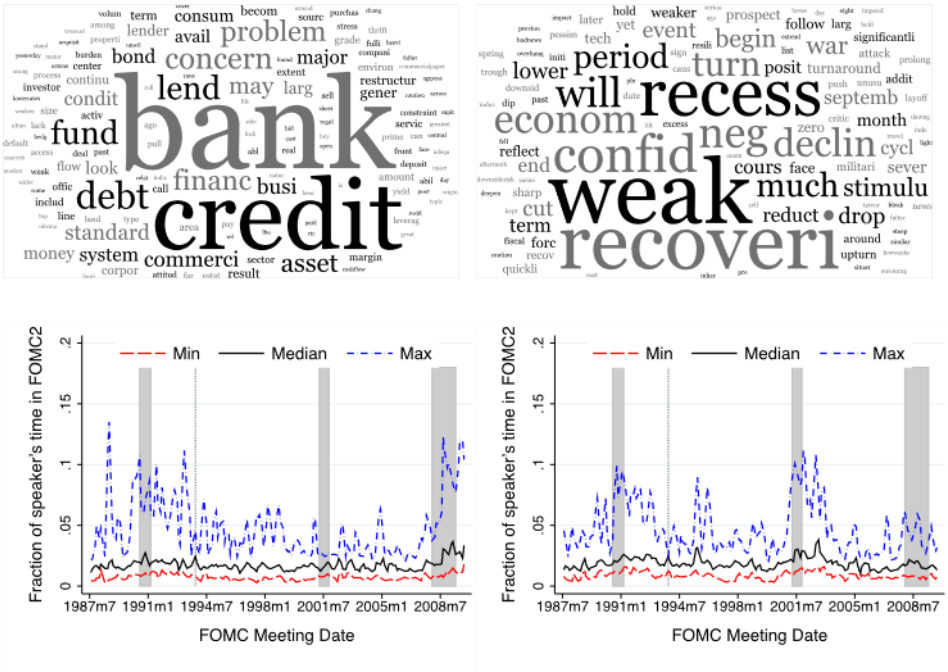
Wallach, H., Mimno, D., and McCallum, A. (2009). Rethinking LDA: Why Priors Matter. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Widmer, P., Galletta, S., and Ash, E. (2020). Media slant is contagious. *Center for Law & Economics Working Paper Series*, 14.

Xu, F. F., Alon, U., Neubig, G., and Hellendoorn, V. J. (2022). A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, pages 1–10.

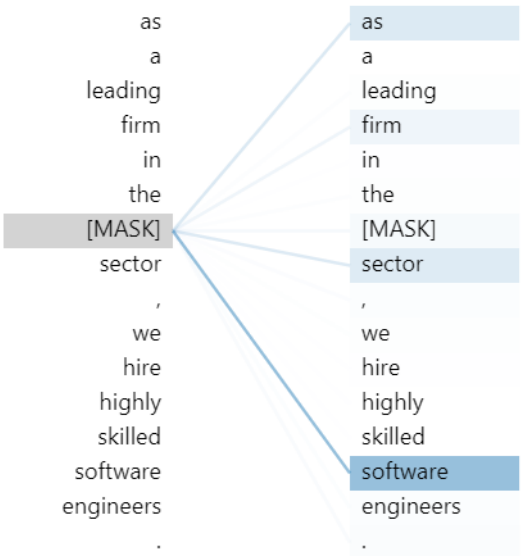
Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al. (2020). Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.



(A) Topic 38 'Financial sector' (B) Topic 39 'Economic weakness'

Figure 1: Illustration of Output of Latent Dirichlet Allocation

This figure, from Hansen et al. (2018), illustrates the output of LDA estimated on the corpus of Federal Open Market Committee transcripts. The word clouds represent two topic-term distributions $\boldsymbol{\beta}_{38}$ and $\boldsymbol{\beta}_{39}$. The size of the words is approximately proportional to the frequency of the term in the topic. Document-topic distributions $\boldsymbol{\theta}_{it}$ are estimated for each FOMC member i and meeting t. The time series plots at the bottom show the maximum, median, and minimum value of these distributions for each meeting across members.



Word	Probability
IT	8.0%
Automotive	7.9%
Technology	7.2%
Healthcare	5.8%
Insurance	5.3%

Figure 2: Example of Self Attention for Masked Word Prediction

This figure illustrates the attention weights estimated for masked word prediction in a corpus of English-language online job postings provided by Lightcast (formerly Emsi Burning Glass). The Transfomer model estimated for this task is DistilBERT (Sanh et al. 2020). In the figure on the left, words in the right column have a darker shading when they are given more weight in the prediction problem. The table on the right lists the most likely words underlying the [MASK] token estimated by the model. See Hansen et al. (2022) for more details.

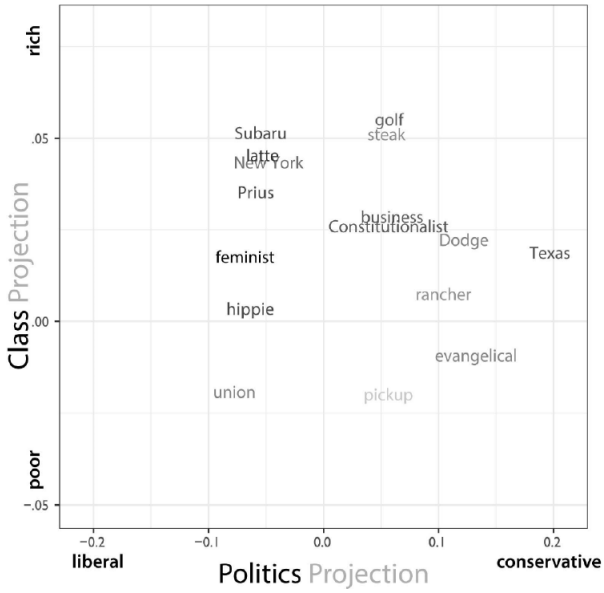
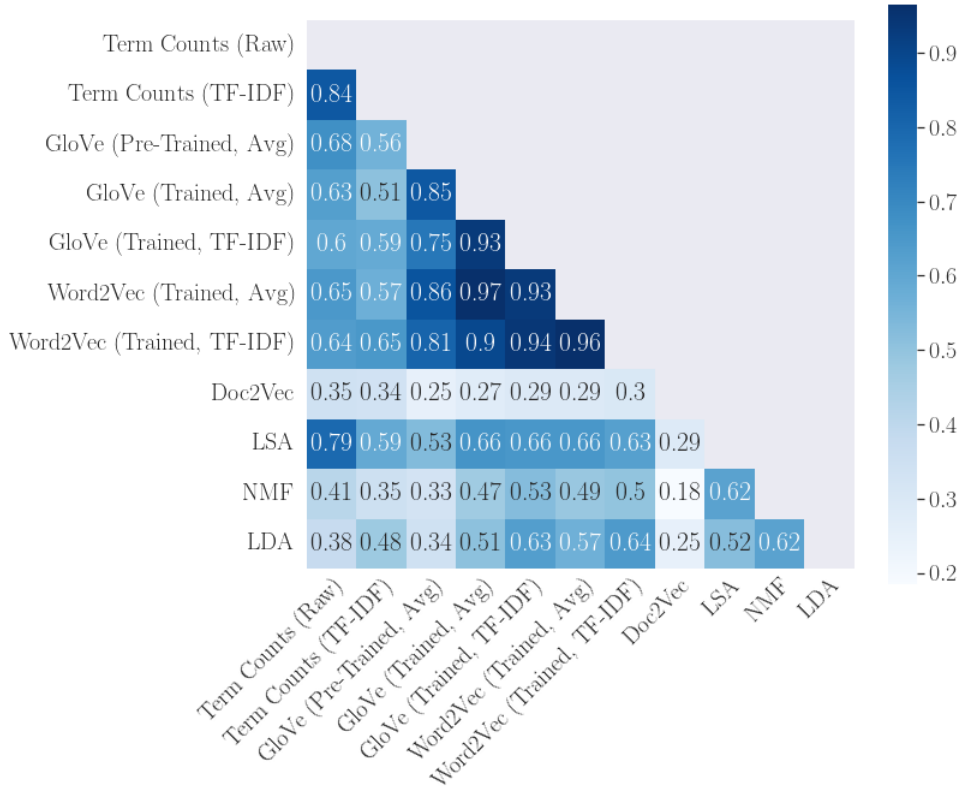
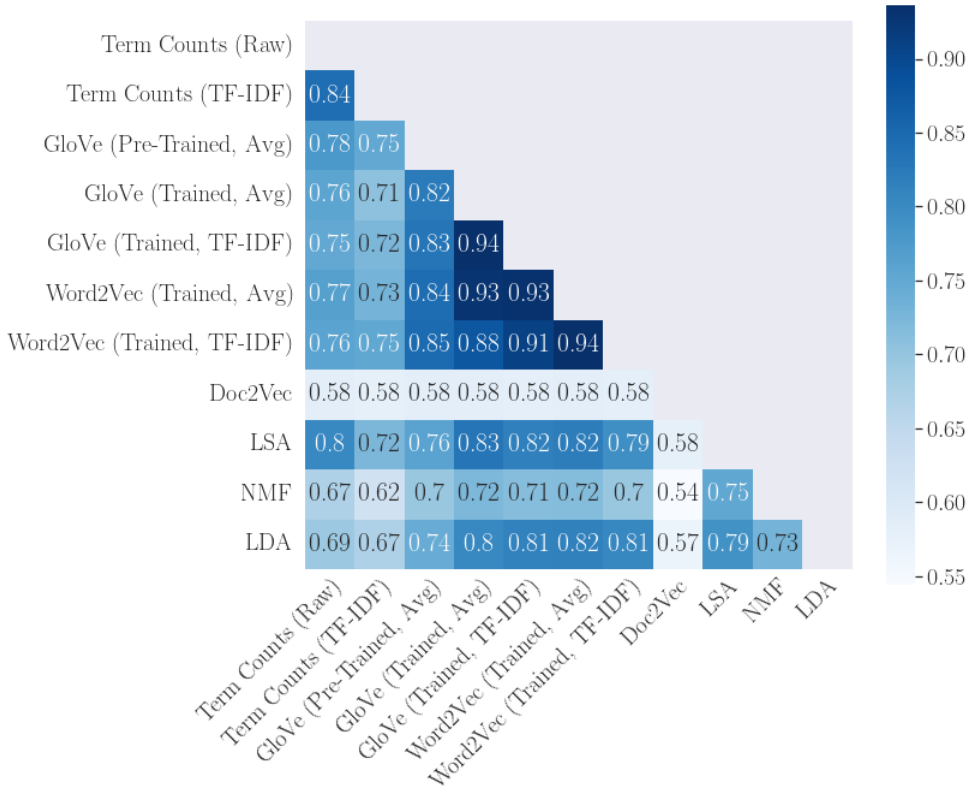


Figure 3: Location of Terms in Class and Politics Attribute Spaces

This figure, taken from Kozlowski et al. (2019), illustrate how words embeddings can be used to associate terms with concepts. The location of a term on the horizontal axis reflects its relative similarity to a set A of words associated with conservative political stances and another set B associated with liberal stances. The further a term is to the right, the closer it lies to A relative to B. Similarly, the position of a term on the vertical axis is related to a poor-rich scale defined by other word sets.



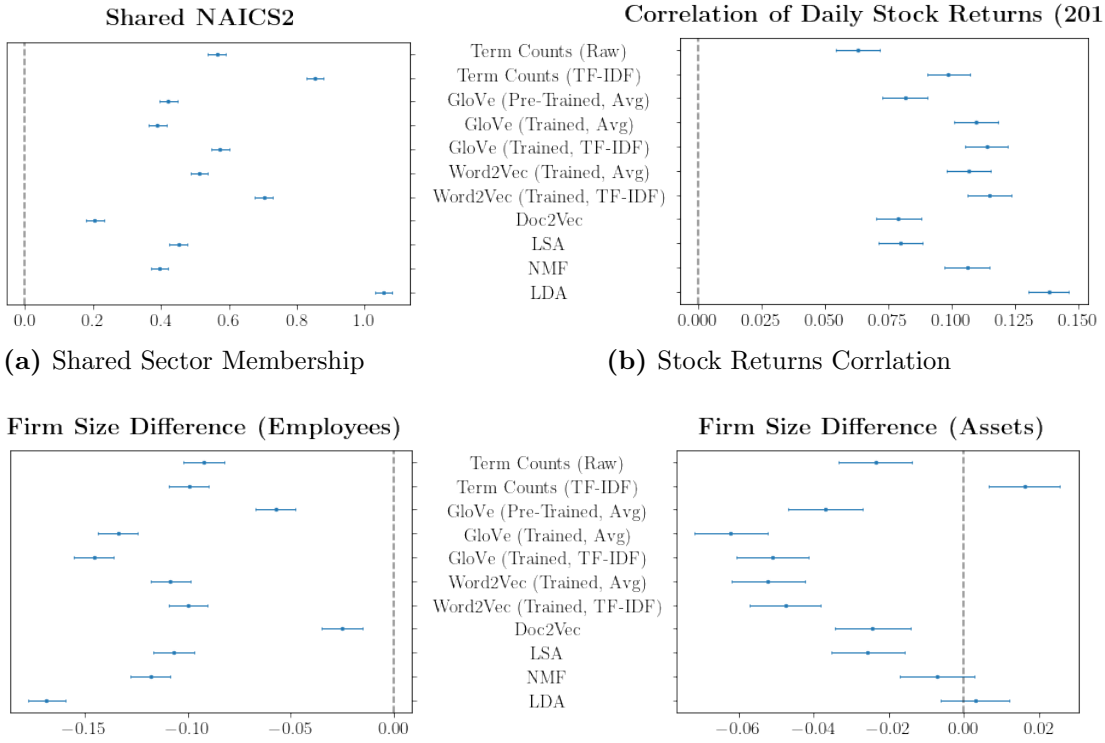
(a) Pearson Correlation for Pairwise Similarity



(b) Agreement Rates for Pairwise Ranking

Figure 4: Comparison of Algorithms for Measuring Document Similarity

We begin with the corpus of *Risk Factors* sections of firms' 2019 10-K filings and compute pairwise cosine similarities across firms according to each of eleven different algorithms. The top panel presents the Pearson correlation between similarity scores produced by each pair of algorithms. For the bottom panel, we draw 10,000 random document triplets, and for each triplet and algorithm, we record whether the second or third document is closest to the first. The bottom panel presents agreement rates between algorithms in this ranking exercise. Two algorithms that produce independent rankings will agree in half of cases, so the scale varies from 0.5 to 1.



(a) Shared Sector Membership

(b) Stock Returns Corrlation

(c) Log Employent Ratio (Abs. Value)

(d) Log Asset Ratio (Abs. Value)

Figure 5: Impact of Algorithm on Downstream Regression Coefficient Estimates

We draw 50,000 random pairs of firms among the population for which we can retrieve 1) a 2019 *Risk Factors* section and 2) a stock price for every trading day in 2019 from CRSP. For each algorithm, we then compute the pairwise similarity between each firm's texts and regress it on a dummy variable for shared NAICS2 sector; the correlation between daily returns in 2019; the absolute log ratio of employees; and the absolute log ratio of total assets. The data on sector and firm size come from COMPUSTAT. The panels in the figure display the point estimates and 95% confidence intervals for each regression coefficient and each algorithm. In all regressions, continuous covariates are expressed in standard deviation units.