ANNUAL REVIEWS

Annual Review of Economics

Text Algorithms in Economics

Elliott Ash[1] and Stephen Hansen[2,3]

[1]Center for Law and Economics, ETH Zurich, Zurich, Switzerland

[2]Department of Economics, University College London, London, United Kingdom; email: stephen.hansen@ucl.ac.uk

[3]Centre for Economic Policy Research, London, United Kingdom

Annu. Rev. Econ. 2023. 15:659–88

First published as a Review in Advance on July 5, 2023

The *Annual Review of Economics* is online at economics.annualreviews.org

https://doi.org/10.1146/annurev-economics-082222-074352

JEL codes: C18, C45, C55

Keywords

text as data, topic models, word embeddings, large language models, transformer models

Abstract

This article provides an overview of the methods used for algorithmic text analysis in economics, with a focus on three key contributions. First, we introduce 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, we define four core empirical tasks that encompass most text-as-data research in economics and enumerate the various approaches that have been taken so far to accomplish these tasks. Finally, we flag limitations in the current literature, with a focus on the challenge of validating algorithmic output.

1. INTRODUCTION

Text analysis in economics is not new. Classic examples include the works of Coase (1960), who analyzed legal cases to study how the law resolves externality problems, and of Friedman & Schwartz (1963), who pioneered the construction of policy surprises via historical documents.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 become 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 a common 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 her *class*."

To illustrate the implementation and application of these algorithms, we build and refer to a companion GitHub page (**https://github.com/sekhansen/text_algorithms_econ**) with reusable code for teaching and research purposes. The examples use publicly available data and source code so that readers can replicate our results and extend them to other data sets. 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: (*a*) measuring the similarity among documents; (*b*) measuring economic concepts contained in raw text; (*c*) measuring how concepts are related to each other in text; and (*d*) 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 to 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

[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).

relevance than computer scientists, so merely borrowing validation tasks from other fields is insufficient. To illustrate the problem, we adapt 10 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 broadly, we advocate for the development of generic, economically relevant language tasks to enable researchers to discriminate among models.

In general, 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**. Here we do not treat the question of how to specify model parameters. Doing so requires some selection criteria, which is part of the larger issue of how to design validation tasks for economic research that we discuss in 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 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. For example, linguistic parsing algorithms for determining grammatical structure typically operate at the sentence level, whereas 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 elsewhere (e.g., Manning et al. 2008, Grimmer & Stewart 2013, Denny & Spirling 2018, Gentzkow et al. 2019a). The basic steps are tokenizing (splitting on whitespace/punctuation), dropping nonletter characters,

[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) and 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).

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 preprocessed 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 preprocessing 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 algorithms (e.g., Jurafsky & Martin 2020, Ash et al. 2023). 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. The process begins by assigning to each unique vocabulary term a unique index value from the integers $1, \dots, V$ 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, $\mathbf{x}_d = (x_{d,1}, \dots, x_{d,V})$ be the vector of counts, and $\boldsymbol{X}$ be the document-term matrix formed by stacking the $\mathbf{x}_d$ across rows.

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

2.3. Dimensionality Reduction

Often we do not care about the particular words documents use but rather 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 that reflects more relevant heterogeneity than the high-dimensional term space.

For economists, one of the most familiar dimensionality reduction techniques is principal component 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 data sets represent deeper structure.

[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. It does not necessarily produce English words, and in some cases it 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. Readers are referred to the companion GitHub repository for additional details on preprocessing.

[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 that can hinder interpretation of its outputs. The statistics literature has linked PCA and factor models for Gaussian distributions (Tipping & 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 statistical 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 $\theta_d \in \Delta^{K-1}$. 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

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

where $\boldsymbol{\Theta}$ is a $(D \times K)$ row-stochastic matrix and $\boldsymbol{B}$ 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 $\boldsymbol{X}$ to a term-frequency matrix $\boldsymbol{X}'$—i.e., we divide each row d by the document length N_d—we can view maximization of Equation 2 as finding the $\boldsymbol{\Theta}$ and $\boldsymbol{B}$ that best approximate $\boldsymbol{X}'$. That optimization objective is also known as nonnegative matrix factorization (NMF).

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

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 the Federal Open Market Committee (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

The vector $\mathbf{x}_d$ 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

[6]The original LDA paper (Blei et al. 2003) only placed a Dirichlet prior on θ_d terms and allowed it to be nonsymmetric. 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 θ_d, and $50/K$, as suggested by Griffiths & Steyvers (2004). Readers are referred to Wallach et al. (2009) for additional thoughts on prior selection in LDA.

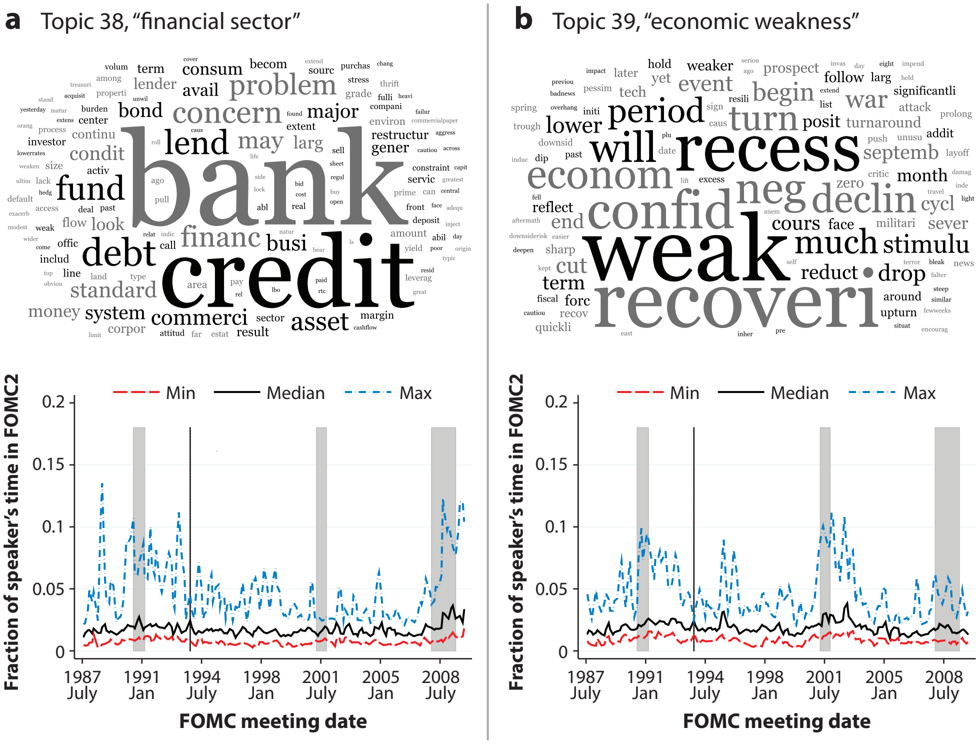

Figure 1

Illustration of output of latent Dirichlet allocation estimated on the corpus of Federal Open Market Committee (FOMC) transcripts. FOMC2 is the part of the meeting in which policy makers deliberate on the correct policy action. The word clouds represent the two topic-term distributions (*a*) β_{38} and (*b*) β_{39}. The size of the words is approximately proportional to the frequency of the term in the topic. Document-topic distributions θ_{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. Figure adapted with permission from Hansen et al. (2018).

n-grams, in practice the size of the feature space V quickly explodes in n. More subtly, a particular word's meaning may depend not only on its immediate neighbors but also on longer-range local dependencies within the 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 (standing for "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 $C(w_{d,n}) = (w_{d,n-L}, \ldots, w_{d,n-1}, w_{d,n+1}, w_{d,n+L})$, that is, a length-$2L$ window surrounding $w_{d,n}$. Then define $\boldsymbol{W}$ as the $V \times V$ word co-occurrence matrix, where an item W_{ij} is the number

[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.

of times that word i appears within an L-tokens span of j (and vice versa; hence $\boldsymbol{W}$ 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

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

where $f(\cdot)$ is a nonnegative, increasing, and concave weighting function.[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 regularly co-occur tend to have vectors with a high dot product. [In Pennington et al.'s (2014) work, the objective also includes non-interacting bias vectors for each term, which we suppress here for conceptual clarity.]

An equally influential word embedding model is Word2Vec (Mikolov et al. 2013a,b; see also Bengio et al. 2003), 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]

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

Word2Vec learns word vectors and context vectors to maximize the 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]

---

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

[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 $\boldsymbol{B}$ from Equation 2 contains a series of row vectors corresponding to each term in the vocabulary (see also Levy & Goldberg 2014). Those vectors contain information about word co-occurrence at the document level, rather than within a local context.

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 data sets, though, there is not enough information to learn reliable vectors. In these cases, one can use pretrained 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 underexplored 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:

1. As a leading firm in the [MASK] sector, we hire highly skilled software engineers.
2. 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 sentence 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 that 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

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

That is, each embedding in the transformed sequence is itself a weighted average of the embeddings in the initial sequence. The nonnegative 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. In modern language models, attention weights are estimated (along with other model parameters) to successfully perform masked-word prediction or other language-related prediction tasks.

Besides these gains in capturing contextual semantic information, a major advantage of attention functions is that massive neural networks composed of stacked attention and feedforward neural network layers (i.e., Transformers) can be efficiently parallelized for training using specialized processors.[15] Beginning with BERT (Devlin et al. 2019), Transformer-based, pretrained language models have consistently set new performance standards for NLP tasks and in the

[13]For an illustration of this strategy, readers are referred to 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 optical character recognition errors.

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

Table 1 Predictions for masked words in example sentences

"Software engineers" sentence		"Petroleum engineers" sentence	
Word	**Probability**	**Word**	**Probability**
it	0.08	energy	0.279
automotive	0.079	oil	0.27
technology	0.072	petroleum	0.088
healthcare	0.058	mining	0.035
insurance	0.053	defense	0.021
software	0.041	automotive	0.02
engineering	0.031	construction	0.017
public	0.03	gas	0.017
infrastructure	0.028	engineering	0.016
financial	0.028	water	0.012

This table displays masked word prediction probabilities for the two example sentences in the main text. The training corpus for estimating these probabilities is English-language online job postings provided by Lightcast (formerly Emsi Burning Glass). The Transformer model used for the task is DistilBERT (Sanh et al. 2020). Readers are referred to Hansen et al. (2023) for more details.

process have become enormously influential. Further well-known models include RoBERTa (Liu et al. 2019), PALM (Chowdhery et al. 2022), and the GPT family (Radford et al. 2018, Brown et al. 2020, OpenAI 2023). Applying a self-supervised approach like Word2Vec, these models are pretrained 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 175 billion estimated parameters, and more recent models are considerably larger.

Table 1 shows the most likely masked words for the two example sentences above produced by a particular Transformer model (Sanh et al. 2020). Although the sentences differ only in one word, which lies several tokens away from the masked word, the model produces distinct predictions that reflect how even seemingly small changes in context produce large differences in meaning. 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.

While this size and complexity of Transformer models have 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 pretrained 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 these are 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-based 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 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 nonlinearities 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 have only a few hundred. Ng & Jordan (2001) argue that joint models $p(y_d, \mathbf{x}_d)$ have worse asymptotic prediction error than conditional models $p(y_d|\mathbf{x}_d)$ but reach their asymptotic limit faster. Hence, data sets with relatively few observations might benefit from modeling this additional structure. Some example methods in this vein are supervised LDA (McAuliffe & 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, pretrained 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. Such an approach will usually dominate bag-of-word models and can approach human performance. Further, because the pretrained Transformer models have a quite general understanding of diverse texts, fine-tuning can achieve good performance even with relatively few labeled training samples.

One restriction of Transformer-based models is limited interpretability, which we discuss further in Section 5. Another is that, until recently, they operated only on relatively short documents. This works well for sentences or paragraphs but not for longer documents such as political speeches, judicial opinions, or corporate filings. While latest-generation Transformer models take in longer inputs (e.g., Beltagy et al. 2020, Zaheer et al. 2020), it can be better with long documents to use non-Transformer-based alternatives such as gradient boosting applied to $\mathbf{x}_d$ (as mentioned above).[16]

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 & 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.

[16]Another option for long documents is the model by 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.

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. This value is higher when the angle between two vectors is smaller, that is, when they share similar directions in the vector space. This metric ensures that similarity is driven by similar word use rather than by document length, as would be the case with Euclidean distance.

These metrics for distance can be used not only for making pairwise document comparisons but also for forming clusters of related documents. A popular method for clustering is k-means, which initializes cluster centroids randomly and then iteratively assigns each document to the nearest centroid, updates the centroid to the mean of the documents in its cluster, and repeats until convergence. The number of clusters, k, is a hyperparameter that needs to be chosen beforehand based on the application. An advantage of clustering, relative to topic models, is that it works on arbitrary vector representations of documents (rather than being limited to term counts, as in LDA). Further, documents are tied to a single cluster rather than having a distribution over multiple topics. Hoberg & Phillips (2016) use a clustering method applied to product descriptions to construct industry categories. Ash et al. (2023) apply clustering to embeddings of entity phrases, which works to dimension-reduce the set of entities mentioned in a corpus and interpretably identify connections among them.

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

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

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 the works of Cagé et al. (2020), who use the distance between online news articles and social media posts to group items into common stories; of Kelly et al. (2021b), who analyze the novelty and influence of technologies using pairwise comparisons between US patent filings; and of Biasi & 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 US federal rule-making process. Hansen et al. (2018) use LDA applied to US Federal Open Market Committee transcripts to measure policy makers' herding behavior following an increase in transparency.

Another approach uses word embeddings to represent documents. In this case, the vector for document d is $\frac{1}{N_d}\sum_n \hat{\boldsymbol{\rho}}_{w_{d,n}}$, that is, 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.

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 (EPU; Baker et al. 2016), skill demand in the labor force (Deming & Kahn 2018), economic sentiment (Shapiro et al. 2022), 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 value terms built by social psychologists (the Moral Foundations Dictionary) to analyze a communal-versus-universalist dimension in congressional speeches.[18] Hassan et al. (2019) build dictionaries of political language based on phrases' simultaneous presence in political science textbooks and absence in general financial language.[19] Second, one can use domain expertise to build term sets from scratch, such as the financial sentiment dictionaries of Loughran & 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 to, 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*.[20] Further, syntactic dependency tags identify the connections between words—for example, which noun is the subject and which is the object.[21] Ash et al. (2020c) extract syntactic dependencies from labor union contracts to extract modal verbs (e.g., *shall*, *may*) that 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.

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.

[17]For example, one can match on a binary indicator 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 term sets to detect the presence of EPU in individual newspaper articles: a set of economic terms, a set of uncertainty terms, and a set of policy-related terms. Newspaper articles are tagged as containing EPU language if they contain a term from each set.

[18]The Moral Foundations Dictionary is available at **https://moralfoundations.org/**.

[19]Similarly, Mastrorocco & 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.

[20]Part-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 & Martin 2020).

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

3.2.2.1. 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, for example, the literature on central bank communication: Boukus & Rosenberg (2006) use LSA and Hansen & 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 and interpreted the content of topics in terms of economic phenomena (Mueller & Rauh 2018, Larsen & 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 & Spirling 2018).

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.

3.2.2.2. Dictionaries augmented with machine judgment. Another approach is to specify an initial set of seed words that reflect a concept and then 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 retrieve either the nearest neighbors of the average embedding vector or of 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 & Hoberg 2019, Atalay et al. 2020, Davis et al. 2020, Li et al. 2021, Soto 2021). In political economy, Gennaro & Ash (2022) use this method to populate emotionality and rationality term sets and also to remove words that are outliers according to cosine similarity. Truffa & 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 river bank. Word embeddings—especially pretrained embeddings learned from generic 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 à 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.

[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.

3.2.2.3. 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 & Ash (2022) 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 US Congress members.

3.2.2.4. Machine prediction based on human annotations. The most accurate approach to concept detection is perhaps direct human reading with appropriate domain expertise. However, 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. 2022). 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 done by Baker et al. (2016).

Hansen et al. (2023) compare 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.

However, only the second case 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 & 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 that terms from each dictionary co-occur within a local window (Apel & Blix Grimaldi 2014, Hassan et al. 2019, Cieslak et al. 2021). Relatedly, Byrne et al. (2023a,b) use pattern-matching techniques to associate concepts to a temporal dimension reflecting the past, present, or future.

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 & 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 the opposite 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 2** (from Kozlowski et al. 2019) locates various terms in two separate conceptual dimensions built with a US-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 by Ash et al. (2020b), who measure gender attitudes of individual US appellate court judges by applying WEAT separately to each judge's authored opinions and considering the correlation between male-female and career-family dimensions. The gender attitudes of judges relate to their decisions and treatment of female colleagues.[24]

These word embedding–based measurements of connections between concepts are based on local co-occurrence of words. This approach misses potentially important context in how the concepts are related; for example, the connection between *driver* and *policeman* is the same in all three of these text snippets: "the policeman killed the driver," "the policeman did not kill the driver," and "the driver killed the policeman." These types of actions and relationships and their directionality—that is, who does what to whom—are key to narratives and to human expression. Ash et al. (2023) show how to use linguistic annotations for agents (the actors) and patients (the targets of actions) to construct and quantify such directed connections. In their application to US congressional speeches, the resulting micro narratives can be fit together in a network to represent distinctive partisan worldviews of US legislators.

3.4. Problem IV: Associating Text with Metadata

In some situations, text comes with metadata that form 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 offered by Gentzkow & Shapiro (2010), who use the political party associated with speakers in the US 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

[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.

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

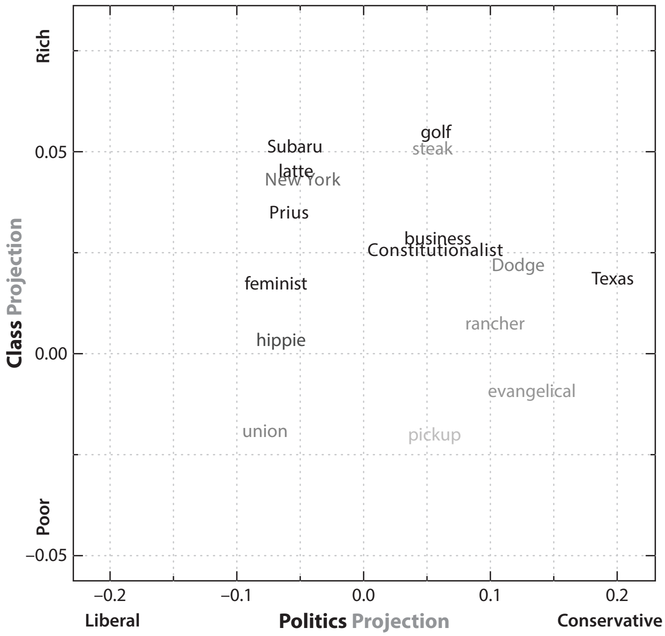

Figure 2

Location of terms in class and politics attribute spaces. This figure illustrates how word 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. Figure adapted with permission from Kozlowski et al. (2019).

of supervised transfer learning. Similarly, Widmer et al. (2020) produce a measure of slant based on the similarity of newspaper language to that used by Fox News.

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 to ensure that the unlabeled documents have the same association between words and outcomes as the training corpus. Osnabrügge et al. (2023) 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.

4. TEXT MEASURES AND ECONOMETRIC MODELS

Each of the four measurement problems outlined above convert text into a quantitative measure. The papers we discussed 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 vector autoregressive model 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 that of Fox News than of other cable news networks.

For the most part, text quantification algorithms and econometric models are treated separately, where the former create data that are 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.

Statistical models of text provide a means of specifying a joint distribution over words and covariates that can be used for valid inference. For example, Taddy (2013, 2015) models the multinomial probability $\mathbf{q}_d$ in Equation 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 θ_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 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 joint models do not address more fundamental issues of identification and, more specifically, of nonclassical 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

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

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 (Section 5.1) and interpretability (Section 5.2) in the use of text algorithms. The section concludes by assessing the prospects of large pretrained language models (Section 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 work 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 & 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. Preprocessing of the documents and further implementation details are described in the **Supplemental Materials**. To compare similarity, we use 10 different approaches to construct document vectors, all of which have appeared in the literature:

Supplemental Material >

- Bag-of-words-based term counts: (*a*) raw counts and (*b*) TF-IDF-weighted term counts.
- Average word embeddings based on (*c*) pretrained GloVe (pretrained on Wikipedia), (*d*) GloVe estimated on the risk factors corpus, (*e*) same as *d* but using TF-IDF weights to compute average, (*f*) Word2Vec estimated on the risk factors corpus, and (*g*) same as *f* but using TF-IDF weights to compute average.
- Dimensionality reduction of document-term matrix: (*h*) LSA, (*i*) NMF, and (*j*) LDA.

We first compute pairwise cosine similarities across firms according to each method. **Figure 3*a*** 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 the third document is closest to the first. **Figure 3*b*** 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 **Figure 3*a*** is 0.64, while the average agreement rate from **Figure 3*b*** is 0.78 (where independent rankings produce an agreement rate of 0.5).

In the **Supplemental Materials**, we describe a similar exercise for word comparisons using four different word embeddings models. **Supplemental Figure B.2** is the analogue of **Figure 3** 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 for 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 pairs of firms and again compute cosine similarities using each method. We then regress each set of pairwise

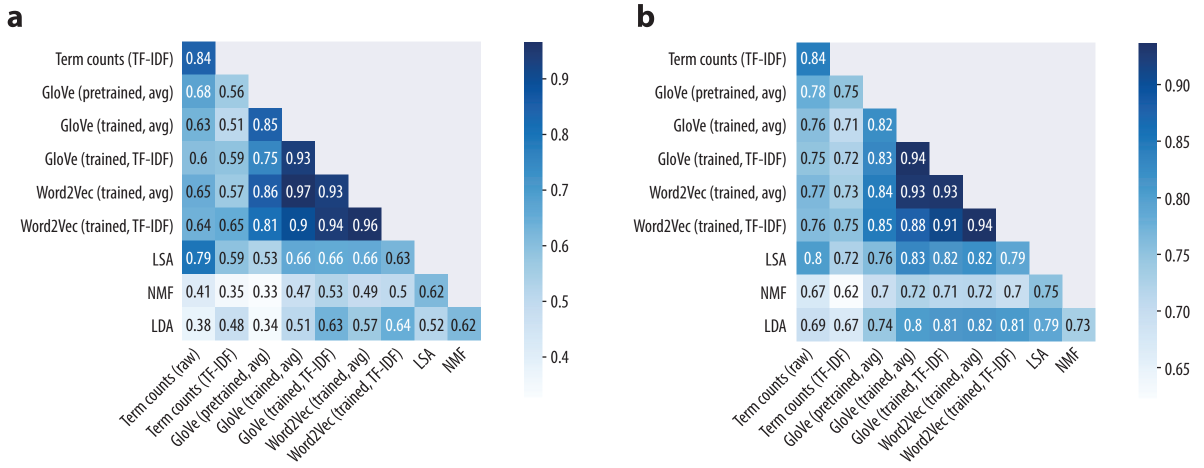

Figure 3

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 10 different algorithms. Panel *a* presents the Pearson correlation between similarity scores produced by each pair of algorithms. For panel *b*, 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. Panel *b* presents agreement rates between algorithms in this ranking exercise. Two algorithms that produce independent rankings will agree in half of the cases, so the scale varies from 0.5 to 1. Abbreviations: LDA, latent Dirichlet allocation; LSA, latent semantic analysis; NMF, nonnegative matrix factorization; TF-IDF, term frequency-inverse document frequency.

similarities on a set of covariates comparing the firms based on whether the firms share a NAICS2 (North American Industry Classification System 2) 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 4 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. However, 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

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

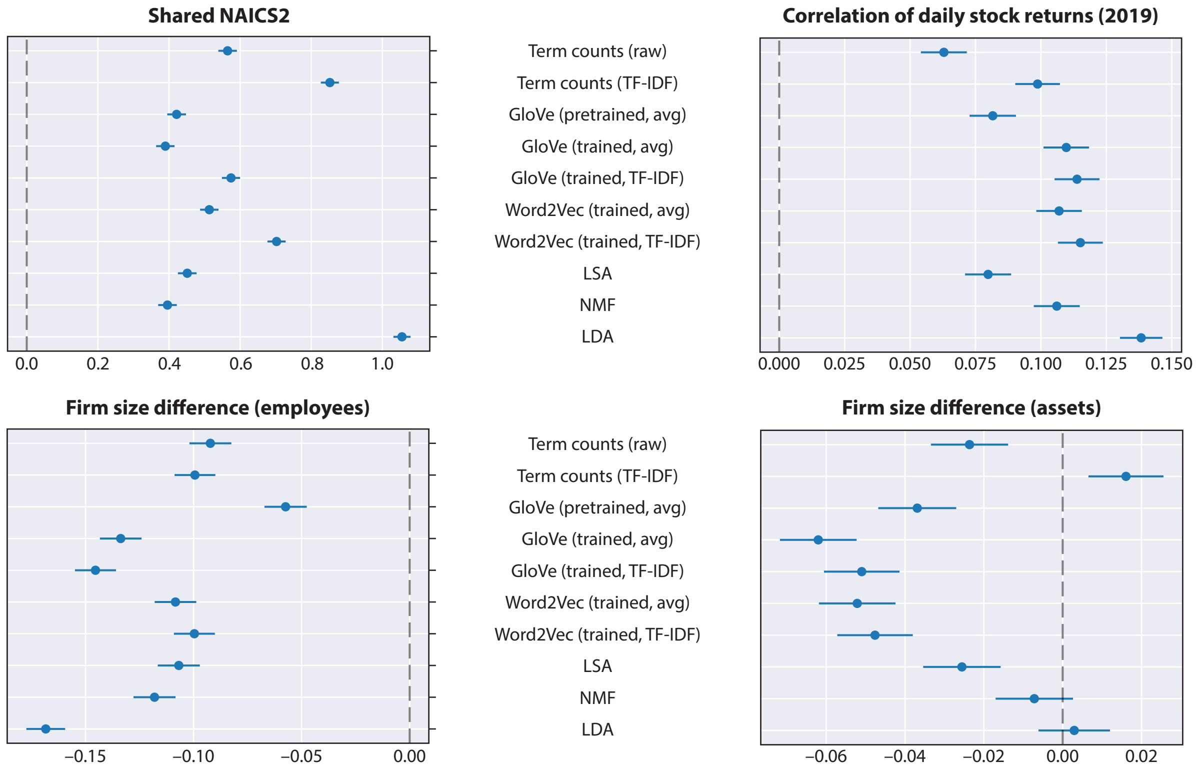

Figure 4

Impact of algorithm on downstream regression coefficient estimates. We draw 50,000 random pairs of firms among the population for which we can retrieve a 2019 Risk Factors section and a stock price for every trading day in 2019 from the Center for Research in Security Prices. For each algorithm, we then compute the pairwise similarity between each firm's texts, which we regress 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. Abbreviations: LDA, latent Dirichlet allocation; LSA, latent semantic analysis; NAICS2, North American Industry Classification System 2; NMF, nonnegative matrix factorization; TF-IDF, term frequency-inverse document frequency.

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 & Ash (2022) 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

[27] Rodriguez & Spirling (2022) conduct this exercise in a political science context.

filling [MASK] with "output."[28] Such basic annotations based 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 10 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. However, allowing each study 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.[29] Whether and how such ex ante validation baselines could be developed, and whether they would actually work in practice, remain to be seen.

A number of other recent technical advances can facilitate the development of these validation baselines. First, the human annotation process can be sped up by machine support—for example, by active learning procedures in which 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 is 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 pretrained language models like GPT-4 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 a held-out sample. However, this choice may not be the most interpretable: Attention-based, deep neural networks 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 US context) would relate to

[28] Generic pretrained 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*, *football*, and *coach*. In economics, the word *team* is used more often in discussions of production and firm organization than in discussions of sports.

[29] Ahrens & McMahon (2023) take initial steps in this direction for monetary economics.

small government, the importance of religion, immigration restrictions, etc. The issue is that *Texas* is likely to co-occur with such terms and to 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 that 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 the 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 are 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 & 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 they do not immediately deliver a solution. The simplest response is then to preprocess documents to remove those correlated features, but the consequences of such targeted preprocessing 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 that are less and less transparent (e.g., as of writing, OpenAI has not disclosed the specific model architecture nor the training data underlying GPT-4), this issue is arguably even more salient than for other machine learning approaches in economics.

5.3. Possibilities of Large Pretrained Language Models

On a more forward-looking note, it is worth revisiting the intriguing and powerful properties of pretrained language models such as BERT and GPT-4. 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 that pretrained language models can help overcome. High-performing machine translation systems are now available as open-source packages (e.g., Tiedemann & 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 & Schwenk 2019).

More speculatively, to the extent that models like those in the GPT family understand language, they may be able to take over language-related research tasks in economics. As an example, consider this article'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."[30] The authors have used GPT models in other similar tasks, such as generating paper titles. In an accompanying notebook we illustrate how to use such models to solve language tasks.[31] In the newest generation of models, in which the generation process is further steered based on human feedback, the performance on these tasks has continued to significantly improve and approach a more generalized artificial intelligence (Ouyang et al. 2022, Bubeck et al. 2023).

Beyond support for writing papers, language models can aid in software development (e.g., Xu et al. 2022) and will likely support code development for economics research, including data wrangling and regression analysis. These models will be able to generate well-formatted tables and other means of reporting results. 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, large language models will be useful in data labeling and validation. So, for example, labeling documents as belonging to a category, or having some feature, should be possible with GPT-like models, perhaps with additional human supervision (e.g., Gilardi et al. 2023, Hansen et al. 2023). 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. In order to make most effective use of large language models, researchers must take care to design appropriate prompts. The sidebar titled GPT-4's Evolving Classification in Response to Repeated Questioning records the interaction between a human and GPT-4 in attempting to solve a classification problem. The model's output is sensitive to specific wording and human-provided hints. The general principles underlying robust prompt design for economics will need to be established to take full advantage of such models' power.

GPT-4'S EVOLVING CLASSIFICATION IN RESPONSE TO REPEATED QUESTIONING

This sidebar reproduces verbatim the interaction between a human and GPT-4. The classification chosen by GPT-4 depends on how the human formulates and organizes prompts.

Human:
You are a data expert working for the Bureau of Labor Statistics specialized in analyzing job postings. Your task is to read the text of fragments of job postings and classify them into one of four categories based on the degree of remote work they allow. Pay close attention to the context in which the words are used, and focus on the actual remote work possibility for the job position rather than the mention of remote work-related terms in other contexts. The four categories and their definitions are:

1. No remote work: The text doesn't offer the possibility of working any day of the week remotely.
2. Hybrid work: The text offers the possibility of working one or more days per week remotely but not the whole week.
3. Fully remote: The text offers the possibility of working all days of the week remotely.
4. Unspecified remote: The text mentions the possibility of working remotely but doesn't clearly specify the extent of this possibility.

You always need to provide a classification. If classification is unclear, say "Could not classify."

(*Continued*)

[30]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.
[31]The notebook is available at **https://github.com/sekhansen/text_algorithms_econ/blob/main/notebooks/7_gpt_demonstration.ipynb**.

(*Continued*)

Please provide the classification and an explanation for your choice in the following format:

- Classification: [Category Number]. [Category Name]
- Explanation: [Explanatory text]

Text of job posting: "This is a Hybrid position so you will be required to carry out 60% Mac and 40% Wintel across the Library & Student Support (LSS)."

GPT-4:
Classification: 2. Hybrid work
Explanation: The text explicitly states that this is a "Hybrid position," which indicates that the job offers the possibility of working some days remotely and some days on-site.

Human:
Are you sure?

GPT-4:
Upon reevaluating the job posting text, I realize my previous response may have been incorrect. Here is my revised classification:
Classification: 4. Unspecified remote
Explanation: The text mentions "Hybrid" work, but it is unclear if this refers to remote work or to the mix of Mac and Wintel responsibilities. The extent of remote work possibilities is not specified, making the classification fall into the "Unspecified remote" category.

Human:
Pay close attention to the meaning of "hybrid" in this context.

GPT-4:
I apologize for the confusion in my previous responses. After a closer examination of the context, I have revised my classification.
Classification: 1. No remote work
Explanation: In this context, "Hybrid" refers to the mix of responsibilities involving 60% Mac and 40% Wintel. There is no mention of remote work possibilities, so it is classified as "No remote work."

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.[32] Fourth, text has almost exclusively been used in reduced-form empirical exercises, but in principle it can also inform structural estimation. Finally, large pretrained language models may be relevant in many research tasks, including labeling data or even helping to write research papers.

The algorithms we discuss in this review, or close variants, are also useful for representing other unstructured data sets beyond text. Bandiera et al. (2020) use LDA to measure leadership styles of CEOs from a detailed time use survey (see also Draca & 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. (2021) and Adukia et al. (2023) 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.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

E.A. and S.H. 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, Áureo de Paula, Prashant Garg, Friedrich Geiecke, Bryan Kelly, Asaf Manela, Claudia Marangon, Michael McMahon, Hannes Mueller, Yabra Muvdi, Szymon Sacher, Carlo Schwarz, Bryan Seegmiller, Dominik Stammbach, Francesca Truffa, and Ashley Wong. Laura Battaglia, Claudia Marangon, and Yabra Muvdi provided excellent research assistance. We thank GPT-3 for writing the abstract.

LITERATURE CITED

Adams-Prassl A, Balgova M, Qian M. 2020. *Flexible work arrangements in low wage jobs: evidence from job vacancy data*. IZA Discuss. Pap. 13691, Inst. Labor Econ., Bonn, Ger.

Adukia A, Eble A, Harrison E, Runesha HB, Szasz T. 2023. What we teach about race and gender: representation in images and text of children's books. *Q. J. Econ.* In press

Advani A, Ash E, Cai D, Rasul I. 2021. *Race-related research in economics and other social sciences*. CEPR Discuss. Pap. 16115, Cent. Econ. Policy Res., London

Ahrens M, McMahon M. 2023. *Natural language processing for monetary economics: a benchmark*. Work. Pap., Univ. Oxford, Oxford, UK

Angelico C, Marcucci J, Miccoli M, Quarta F. 2022. Can we measure inflation expectations using Twitter? *J. Econometr.* 228(2):259–77

Angrist J, Pischke JS. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton Univ. Press

Apel M, Blix Grimaldi M. 2014. How informative are central bank minutes? *Rev. Econ.* 65(1):53–76

Arora S, Liang Y, Ma T. 2016. *A simple but tough-to-beat baseline for sentence embeddings*. Paper presented at the 5th International Conference on Learning Representations, Toulon, France, Apr. 24–26

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

Artetxe M, Schwenk H. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguist.* 7:597–610

Ash E, Chen D, Naidu S. 2020a. *Ideas have consequences: the impact of law and economics on American justice*. Work. Pap. 4, Cent. Law Econ., ETH Zurich, Zurich, Switz.

Ash E, Chen DL, Ornaghi A. 2020b. *Gender attitudes in the judiciary: evidence from U.S. circuit courts.* Work. Pap., Univ. Warwick, Coventry, UK

Ash E, Durante R, Grebenschikova M, Schwarz C. 2021. *Visual stereotypes in news media*. CEPR Discuss. Pap. 16624, Cent. Econ. Policy Res., London

Ash E, Gauthier G, Widmer P. 2023. Relatio: Text semantics capture political and economic narratives. *Political Anal.* In press. **https://doi.org/10.1017/pan.2023.8**

Ash E, Jacobs J, MacLeod B, Naidu S, Stammbach D. 2020c. *Unsupervised extraction of workplace rights and duties from collective bargaining agreements.* Paper presented at the 2nd International Workshop on Mining and Learning in the Legal Domain (MLLD-2020), online, Nov. 17–20

Ash E, Morelli M, Vannoni M. 2020d. *More laws, more growth? Evidence from US states.* Work. Pap. 15, Cent. Law Econ., ETH Zurich, Zurich, Switz.

Atalay E, Phongthiengtham P, Sotelo S, Tannenbaum D. 2020. The evolution of work in the United States. *Am. Econ. J. Appl. Econ.* 12(2):1–34

Bahdanau D, Cho KH, Bengio Y. 2015. *Neural machine translation by jointly learning to align and translate.* Paper presented at the 3rd International Conference on Learning Representations, San Diego, CA, May 7–9

Baker SR, Bloom N, Davis SJ. 2016. Measuring economic policy uncertainty. *Q. J. Econ.* 131(4):1593–636

Bana SH. 2022. *Work2vec: using language models to understand wage premia*. Work. Pap., Stanford Univ., Stanford, CA

Bandiera O, Prat A, Hansen S, Sadun R. 2020. CEO behavior and firm performance. *J. Political Econ.* 128(4):1325–69

Beltagy I, Peters ME, Cohan A. 2020. Longformer: the long-document transformer. arXiv:2004.05150 [cs.CL]

Bengio Y, Ducharme R, Vincent P, Janvin C. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.* 3:1137–55

Bertrand M, Bombardini M, Fisman R, Hackinen B, Trebbi F. 2021. Hall of mirrors: corporate philanthropy and strategic advocacy. *Q. J. Econ.* 136(4):2413–65

Besley T, Fetzer T, Mueller H. 2020. *How big is the media multiplier? Evidence from dyadic news data*. Unpublished manuscript, London Sch. Econ., London

Biasi B, Ma S. 2022. *The education-innovation gap*. NBER Work. Pap. 29853

Blei DM, Ng AY, Jordan MI. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022

Bloom N, Hassan TA, Kalyani A, Lerner J, Tahoun A. 2021. *The diffusion of disruptive technologies*. NBER Work. Pap. 28999

Bojanowski P, Grave E, Joulin A, Mikolov T. 2017. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* 5:135–46

Boukus E, Rosenberg JV. 2006. *The information content of FOMC minutes*. Work. Pap., Fed. Reserve Bank New York, New York

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, et al. 2020. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 33:1877–901

Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, et al. 2023. Sparks of artificial general intelligence: early experiments with GPT-4. arXiv:2303.12712 [cs.CL]

Bybee L, Kelly BT, Manela A, Xiu D. 2021. *Business news and business cycles*. NBER Work. Pap. 29344

Byrne D, Goodhead R, McMahon M, Parle C. 2023a. *Measuring the temporal dimension of text: an application to policymaker speeches*. CEPR Discuss. Pap. 17931, Cent. Econ. Policy Res., London

Byrne D, Goodhead R, McMahon M, Parle C. 2023b. *The Central Bank crystal ball: temporal information in monetary policy communication.* CEPR Discuss. Pap. 17930, Cent. Econ. Policy Res., London

Cagé J, Hervé N, Viaud ML. 2020. The production of information in an online world. *Rev. Econ. Stud.* 87(5):2126–64

Caliskan A, Bryson JJ, Narayanan A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–86

Chang J, Gerrish S, Wang C, Boyd-Graber J, Blei D. 2009. Reading tea leaves: how humans interpret topic models. *Adv. Neural Inf. Process. Syst.* 22. **https://papers.nips.cc/paper_files/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf**

Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, et al. 2022. PaLM: scaling language modeling with pathways. arXiv:2204.02311 [cs.CL]

Cieslak A, Hansen S, McMahon M, Xiao S. 2021. *Policymakers' uncertainty*. Unpublished manuscript, Univ. Coll. London, London

Coase RH. 1960. The problem of social cost. *J. Law Econ.* 3:1–44

Davis SJ, Hansen S, Seminario-Amez C. 2020. *Firm-level risk exposures and stock returns in the wake of COVID-19*. NBER Work. Pap. 27867

Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. 1990. Indexing by latent semantic analysis. *J. Am. Soc. Inform. Sci.* 41(6):391–407

Deming D, Kahn LB. 2018. Skill requirements across firms and labor markets: evidence from job postings for professionals. *J. Labor Econ.* 36(S1):S337–69

Demszky D, Garg N, Voigt R, Zou J, Gentzkow M, et al. 2019. Analyzing polarization in social media: method and application to tweets on 21 mass shootings. arXiv:1904.01596 [cs.CL]

Denny MJ, Spirling A. 2018. Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Political Anal.* 26(2):168–89

Devlin J, Chang MW, Lee K, 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*, Vol. 1, pp. 4171–86. Stroudsburg, PA: ACL

Ding C, Li T, 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*, Vol. 1, pp. 342–47. Boston, MA: AAAI

Djourelova M, Durante R, Martin G. 2021. *The impact of online competition on local newspapers: evidence from the introduction of Craigslist*. CESifo Work. Pap. 9090, CESifo, Munich, Ger.

Draca M, Schwarz C. 2018. *How polarized are citizens? Measuring ideology from the ground-up*. Warwick Econ. Res. Pap. Ser. 1218, Univ. Warwick, Coventry, UK

Enke B. 2020. Moral values and voting. *J. Political Econ.* 128(10):3679–729

Farrell MH, Liang T, 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. *J. Eur. Econ. Assoc.* 18(6):3337–75

Friedman M, Schwartz AJ. 1963. *A Monetary History of the United States: 1867–1960.* Princeton, NJ: Princeton Univ. Press

Gallagher RJ, Reing K, Kale D, Ver Steeg G. 2017. Anchored correlation explanation: topic modeling with minimal domain knowledge. *Trans. Assoc. Comput. Linguist.* 5:529–42

Gennaro G, Ash E. 2022. Emotion and reason in political language. *Econ. J.* 132(643):1037–59

Gentzkow M, Kelly B, Taddy M. 2019a. Text as data. *J. Econ. Lit.* 57(3):535–74

Gentzkow M, Shapiro JM. 2010. What drives media slant? Evidence from US daily newspapers. *Econometrica* 78(1):35–71

Gentzkow M, Shapiro JM, Taddy M. 2019b. Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica* 87(4):1307–40

Gilardi F, Alizadeh M, Kubli M. 2023. ChatGPT outperforms crowd-workers for text-annotation tasks. arXiv:2303.15056 [cs.CL]

Goldberg Y. 2017. *Neural Network Methods for Natural Language Processing*. San Rafael, CA: Morgan & Claypool

Griffiths TL, Steyvers M. 2004. Finding scientific topics. *PNAS* 101(Suppl. 1):5228–35

Grimmer J, Roberts ME, Stewart BM. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton, NJ: Princeton Univ. Press

Grimmer J, Stewart BM. 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political Anal.* 21(3):267–97

Hanley KW, Hoberg G. 2019. Dynamic interpretation of emerging risks in the financial sector. *Rev. Financ. Stud.* 32(12):4543–603
Hansen S, Lambert PJ, Bloom N, Davis SJ, Sadun R, Taska B. 2023. *Remote work across jobs, companies, and space*. NBER Work. Pap. 31007
Hansen S, McMahon M. 2016. Shocking language: understanding the macroeconomic effects of central bank communication. *J. Int. Econ.* 99:S114–33
Hansen S, McMahon M, Prat A. 2018. Transparency and deliberation within the FOMC: a computational linguistics approach. *Q. J. Econ.* 133(2):801–70
Hansen S, Ramdas T, Sadun R, Fuller J. 2021. *The demand for executive skills*. NBER Work. Pap. 28959
Hassan TA, Hollander S, van Lent L, Tahoun A. 2019. Firm-level political risk: measurement and effects. *Q. J. Econ.* 134(4):2135–202
Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning*. New York: Springer
Hoberg G, Phillips G. 2010. Product market synergies and competition in mergers and acquisitions: a text-based analysis. *Rev. Financ. Stud.* 23(10):3773–811
Hoberg G, Phillips G. 2016. Text-based network industries and endogenous product differentiation. *J. Political Econ.* 124(5):1423–65
Hofmann T. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pp. 289–96. San Francisco, CA: Morgan Kaufmann
Iaria A, Schwarz C, Waldinger F. 2018. Frontier knowledge and scientific production: evidence from the collapse of international science. *Q. J. Econ.* 133(2):927–91
Jha M, Liu H, Manela A. 2022. *Does finance benefit society? A language embedding approach*. Work. Pap., Wash. Univ. St. Louis, St. Louis, MO
Joulin A, Grave E, Bojanowski P, Mikolov T. 2016. Bag of tricks for efficient text classification. arXiv:1607.01759 [cs.CL]
Jurafsky D, Martin JH. 2020. *Speech and Language Processing*. Unpublished manuscript, Stanford Univ., Stanford, CA. 3rd ed. **https://web.stanford.edu/jurafsky/slp3/ed3book.pdf**
Ke S, Olea JLM, Nesbit J. 2021. *Robust machine learning algorithms for text analysis*. Unpublished manuscript, Yale Sch. Manag., Yale Univ., New Haven, CT
Ke ZT, Kelly B, Xiu D. 2019. *Predicting returns with text data*. NBER Work. Pap. 26186
Kelly B, Manela A, Moreira A. 2021a. Text selection. *J. Bus. Econ. Stat.* 39(4):859–79
Kelly B, Papanikolaou D, Seru A, Taddy M. 2021b. Measuring technological innovation over the long run. *Am. Econ. Rev. Insights* 3(3):303–20
Khodak M, Saunshi N, Liang Y, Ma T, Stewart B, Arora S. 2018. A la carte embedding: cheap but effective induction of semantic feature vectors. arXiv:1805.05388 [cs.CL]
Kleinberg J, Mullainathan S. 2019. Simplicity creates inequity: implications for fairness, stereotypes, and interpretability. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp. 807–8. New York: ACM
Kogan L, Papanikolaou D, Schmidt L, Seegmiller B. 2019. *Technology, vintage-specific human capital, and labor displacement: evidence from linking patents with occupations*. NBER Work. Pap. 29552
Kozlowski AC, Taddy M, Evans JA. 2019. The geometry of culture: analyzing the meanings of class through word embeddings. *Am. Sociol. Rev.* 84(5):905–49
Larsen VH, Thorsrud LA. 2019. The value of news for economic developments. *J. Econometr.* 210(1):203–18
Levy O, Goldberg Y. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Vol. 2, pp. 302–8. Stroudsburg, PA: ACL
Li K, Mai F, Shen R, Yan X. 2021. Measuring corporate culture using machine learning. *Rev. Financ. Stud.* 34(7):3265–315
Lippmann Q. 2022. Gender and lawmaking in times of quotas. *J. Public Econ.* 207:104610
Liu Y, Ott M, Goyal N, Du J, Joshi M, et al. 2019. RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692 [cs.CL]
Loughran T, Mcdonald B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J. Finance* 66(1):35–65
Manning CD, Raghavan P, Schütze H. 2008. *Introduction to Information Retrieval*. New York: Cambridge Univ. Press

Mastrorocco N, Ornaghi A. 2020. *Who watches the watchmen? Local news and police behavior in the United States*. Work. Pap., Dep. Econ., Univ. Warwick, Coventry, UK

McAuliffe J, Blei D. 2007. Supervised topic models. *Adv. Neural Inf. Process. Syst.* 20. **https://papers.nips.cc/paper_files/paper/2007/file/d56b9fc4b0f1be8871f5e1c40c0067e7-Paper.pdf**

Mikolov T, Chen K, Corrado G, Dean J. 2013a. Efficient estimation of word representations in vector space. arXiv:1301.3781 [cs]

Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. 2013b. Distributed representations of words and phrases and their compositionality. arXiv:1310.4546 [cs.CL]

Monarch RM. 2021. *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI*. Shelter Island, NY: Manning

Mueller H, Rauh C. 2018. Reading between the lines: prediction of political violence using newspaper text. *Am. Political Sci. Rev.* 112(2):358–75

Ng AY, Jordan MI. 2001. On discriminative versus 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*, pp. 841–48. Cambridge, MA: MIT Press

OpenAI. 2023. GPT-4 technical report. arXiv:2303.08774 [cs.CL]

Osnabrügge M, Ash E, Morelli M. 2023. Cross-domain topic classification for political texts. *Political Anal.* 31(1):59–80

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, et al. 2022. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* 35:27730–44

Pennington J, Socher R, Manning C. 2014. GloVe: global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–43. Stroudsburg, PA: ACL

Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, et al. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 2227–37. Stroudsburg, PA: ACL

Phuong M, Hutter M. 2022. Formal algorithms for transformers. arXiv:2207.09238 [cs.LG]

Pryzant R, Shen K, Jurafsky D, 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*, Vol. 1, pp. 1615–25. Stroudsburg, PA: ACL

Radford A, Narasimhan K, Salimans T, Sutskever I. 2018. *Improving language understanding by generative pre-training*. Work. Pap., OpenAI, San Francisco, CA

Ribeiro MT, Singh S, Guestrin C. 2016. Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 97–101. Stroudsburg, PA: ACL

Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, et al. 2014. Structural topic models for open-ended survey responses. *Am. J. Political Sci.* 58(4):1064–82

Rodriguez PL, Spirling A. 2022. Word embeddings: what works, what doesn't, and how to tell the difference for applied research. *J. Politics* 84(1):101–15

Rudin C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1(5):206–15

Ruiz FJR, Athey S, Blei DM. 2020. SHOPPER: a probabilistic model of consumer choice with substitutes and complements. *Ann. Appl. Stat.* 14(1):1–27

Rydning J. 2021. *Worldwide global datasphere and global storagesphere structured and unstructured data forecast, 2021–2025*. Market Forecast, Int. Data Corp., Needham, MA

Sacher S, Battaglia L, Hansen S. 2021. Hamiltonian Monte Carlo for regression with high-dimensional categorical data. arXiv:2107.08112 [econ.EM]

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

Sedova A, Stephan A, Speranskaya M, Roth B. 2021. Knodle: modular weakly supervised learning with PyTorch. arXiv:2104.11557 [cs.LG]

Shapiro AH, Sudhof M, Wilson DJ. 2022. Measuring news sentiment. *J. Econometr.* 228(2):221–43

Shen Z, Zhang R, Dell M, Lee BCG, Carlson J, Li W. 2021. LayoutParser: a unified toolkit for deep learning based document image analysis. In *International Conference on Document Analysis and Recognition*, pp. 131–46. New York: Springer

Soto PE. 2021. Breaking the Word Bank: measurement and effects of bank level uncertainty. *J. Financ. Serv. Res.* 59(1):1–45

Stammbach D, Antoniak M, Ash E. 2022. Heroes, villains, and victims, and GPT-3–automated extraction of character roles without training data. arXiv:2205.07557 [cs.CL]

Taddy M. 2013. Multinomial inverse regression for text analysis. *J. Am. Stat. Assoc.* 108(503):755–70

Taddy M. 2015. Distributed multinomial regression. *Ann. Appl. Stat.* 9(3):1394–414

Thorsrud LA. 2020. Words are the new numbers: a newsy coincident index of the business cycle. *J. Bus. Econ. Stat.* 38(2):393–409

Tiedemann J, 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*, pp. 479–80. Geneva, Switz.: EAMT

Tipping ME, Bishop CM. 1999. Probabilistic principal component analysis. *J. R. Stat. Soc. B* 61(3):611–22

Truffa F, Wong A. 2022. *Undergraduate gender diversity and direction of scientific research*. Work. Pap., Stanford Univ., Stanford, CA

Vafa K, Naidu S, Blei D. 2020. Text-based ideal points. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5345–57. Stroudsburg, PA: ACL

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. 2017. Attention is all you need. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–10. New York: ACM

Wallach H, Mimno D, McCallum A. 2009. Rethinking LDA: why priors matter. *Adv. Neural Inf. Process. Syst.* 22. **https://papers.nips.cc/paper_files/paper/2009/file/0d0871f0806eae32d30983b62252da50-Paper.pdf**

Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR. 2018. GLUE: a multi-task benchmark and analysis platform for natural language understanding. arXiv:1804.07461 [cs.CL]

Widmer P, Galletta S, Ash E. 2020. *Media slant is contagious.* Work. Pap. 14, Cent. Law Econ., ETH Zurich, Zurich, Switz.

Xu FF, Alon U, Neubig G, Hellendoorn VJ. 2022. A systematic evaluation of large language models of code. In *MAPS 2022: Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*. New York: ACM. **https://doi.org/10.1145/3520312.3534862**

Zaheer M, Guruganesh G, Dubey KA, Ainslie J, Alberti C, et al. 2020. Big Bird: transformers for longer sequences. *Adv. Neural Inf. Process. Syst.* 33:17283–97

Annual Review of Economics

Volume 15, 2023

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Economics* articles may be found at http://www.annualreviews.org/errata/economics