Article

A Tool Kit for Relation Induction in Text Analysis

Sociological Methods & Research
2025, Vol. 54(2) 565–604

DOI: 10.1177/00491241241233242
journals.sagepub.com/home/smr

Dustin S. Stoltz[1], Marshall A. Taylor[2], and Jennifer S. K. Dudley[3]

Abstract

Distances derived from word embeddings can measure a range of gradational relations—similarity, hierarchy, entailment, and stereotype—and can be used at the document- and author-level in ways that overcome some of the limitations of weighted dictionary methods. We provide a comprehensive introduction to using word embeddings for relation induction, and demonstrate how such techniques can complement dictionary methods as unsupervised, deductive methods.

Keywords

computational text analysis, word embeddings, semantic relations, dictionaries, cultural sociology

Social scientists are increasingly turning to computer-assisted text analysis for unsupervised, deductive methods. To do so, analysts commonly use dictionaries methods because they are transparent, reproducible, and computationally efficient. Like all methods, however, they have limitations.

[1]Department of Sociology and Anthropology, Lehigh University, Bethlehem, PA, USA
[2]Department of Sociology, New Mexico State University, Las Cruces, NM, USA
[3]Division of Management, Columbia University Business School, New York, NY, USA

Corresponding Author:
Dustin S. Stoltz, Department of Sociology and Anthropology, Lehigh University, 31 Williams Drive, Bethlehem, PA 18015-3027, USA.
Email: dss219@lehigh.edu

Data Availability Statement included at the end of the article

Here, we focus on using weighted dictionaries as gradational measures at the document- or author-level: specifically, such measures are often strongly associated with the variation in document lengths as a result of the usual distribution of words.

Relation induction offers a technical complement to dictionary approaches as an unsupervised, deductive method. Lexical relations are modeled as vector translations in a word embedding space (Bouraoui, Jameel, and Schockaert, 2018; Mikolov, Yih, and Zweig, 2013a; Pennington, Socher, and Manning, 2014). Word embeddings use term co-occurrence statistics to assign words locations in a multidimensional semantic space. Analysts can easily define a location in that semantic space and "weight" any word in a corpus by its distance from this location. By drawing on a range of post-processing procedures, analysts can measure a variety of relations beyond basic semantic similarity. By then representing documents as locations in that same space, analysts can measure the extent a document engages with a relation by its distance.

In what follows, we (1) briefly present the technical limitations of using dictionary-based graded measures at aggregate (document- or author-) levels. Then, we (2) provide a comprehensive overview of relation induction techniques using word embeddings. We then (3) demonstrate how to use relation induction as gradational measures at the aggregate level.

Limitations of Dictionaries

The *dictionary* is a simple data structure containing pairs of values and keys and is widely used in the social sciences (e.g., Snefjella and Kuperman, 2015; Goldberg et al., 2016; Kovács, Carroll, and Lehman, 2017; Flores, 2017; Frye and Gheihman, 2018; Nelson, 2020; Paxton, Velasco, and Ressler, 2020; Franzosi, 2021; Bhatt, Goldberg, and Srivastava, 2021; Cheng et al., 2023). In computer-assisted text analysis, both qualitative and quantitative, the *key* is usually a unique *string* of characters—keyword, token, term, word, phrase, *n*-gram—which is *matched* and then *tagged* with an associated value. The tagging process can grow more complex with additional rules—such as ignoring case or allowing fuzzy matches with regular expressions—but the basic process remains.

Dictionary methods often produce *categorical* values. This is common in qualitative coding when words and phrases are subsumed under a more encompassing "code" or "theme" (Miles and Huberman, 1994: 44-45; Strauss, 1987: 55-81; Deterding and Waters, 2018; Nelson et al., 2021). Words can also be tagged with a *weight indicating a magnitude within a category*. This kind of dictionary is our focus.

Sentiment analysis, for example, largely relies on an ordinal measure of "polarity" with higher scores indicating a more positive sentiment and lower scores indicating a more negative sentiment. Studying online product reviews (Liu, Hu, and Cheng, 2005), for instance, typically involves either counting the proportion of positive/negative words or assigning words a positivity/negativity rating along a scale (e.g., ranging from +4 to −4) and summing the total. This assigns a *magnitude*, rather than a binary category, to a text.

Dictionaries have limitations. For instance, building and validating them can be very time-consuming and, potentially, idiosyncratic. Using multiple coders, for example, with crowdsourcing platforms, can be less time-intensive and idiosyncratic but is more resource-demanding. However, dictionary methods run into unique problems measuring magnitudes at the document-level (Aslanidis, 2018: 1245-250). Specifically, any method based on counting the number of string matches to measure a *magnitude* faces problems associated with the "long-tail" distribution of words in corpora (Baayen, 2002).

The frequencies of individual words (i.e., tokens) follow a long-tail distribution, where a few words are far more frequent than most others. This holds for every language studied thus far, both natural and synthetic, and most *n*-grams, from sub-word character and phoneme strings to multi-word phrases (Piantadosi, 2014; Yu, Xu, and Liu, 2018). Formally, the frequency rank, r, of a word is inversely proportional to its frequency f—this is Zipf's Law. Most unique words—in many cases, half or more of a corpus—only occur once or twice in a corpus (Kornai, 2008: 72). As the length of a document or corpus increases, the number of distinct tokens (i.e., types) also increases, but at a slowly declining rate of increase—this is the Heaps-Herdan law.

Difficulties result from these two linguistic regularities (see Figure 3 in the supplemental material for an empirical demonstration of both). We might assume that if documents were to increase in size, the proportion of relevant tokens would not change significantly as the *topic* is driving the proportions. But, longer documents have higher chances of including words in our dictionaries than shorter documents (Salton and Buckley, 1988).

Norming schemes attempt to unmoor the number of unique types from the number of tokens. Term frequency-inverse document frequency (*tf-idf*) is a common norming scheme, entailing multiplying a term's relative frequency (i.e., dividing each raw count by the document length) by the (logarithm of the) inverse of the proportion of documents containing the term.[1] Below, we show why norming, unfortunately, only helps so much.

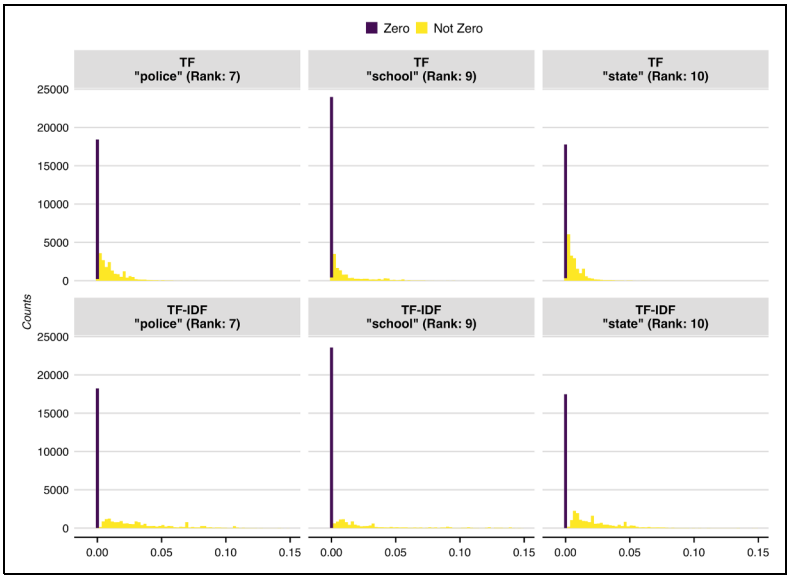

Figure 1. Relative frequencies and term frequency-inverse document frequency (TF-IDF) for highly ranked terms.

Consider the histograms in Figure 1. The *x*-axis is the normed frequency (relative frequency on the top and *tf-idf* on the bottom), and the *y*-axis is the number of times that frequency occurs across the corpus.[2] The data are frequencies from the (preprocessed) *New York Times* articles that make up the Dynamics of Collective Action (DoCA) dataset of protest events, which we describe in more detail in the Appendix in the supplemental material and repeatedly use throughout this article. The terms shown here are very frequent terms.[3] Despite these terms' high frequencies relative to other words in the corpus, they still rarely occur.[4]

Relying on term frequencies alone assumes terms are independent. Weighting terms by the magnitude it indicates a semantic domain relaxes this assumption (Osgood, Suci, and Tannenbaum, 1957). Rather than being treated as independent types (Sidorov et al., 2014), terms that are semantically alike are placed near each other along this dimension. Typically, these weights are then multiplied by the frequency (normed or raw).[5]

Recall that word frequencies are inversely proportional to their rank in a corpus (Zipf's law). As more frequent terms tend to be more generic, they

are also more likely to receive neutral weightings. This means that informative terms, with consequently high ratings, will be infrequent. Further, since the number of distinct words is a linear function of document length (Heaps-Herdan law), we can put these two patterns together to observe that, statistically speaking, *more informative words are rare relative to less informative words, and the informative words we do have are more likely to occur in longer documents*. Dictionary weights can only do so much when the words they are weighting are so infrequent. Furthermore, analysts must disregard or assign low/neutral weight to tokens that are missing from the dictionary—that is, "out-of-vocabulary" (OOV).

Despite the drawbacks of using dictionaries as gradational measures at an aggregate level, the growing number of hand-weighted dictionaries—often painstakingly curated and validated—should not be discarded, nor should other count-based methods (e.g., Wood, 2023). As we will demonstrate, we can use such dictionaries alongside word embedding methods.

What Are Word Embeddings?

Word embeddings use term co-occurrence statistics to summarize the regular features of word usage in writing. Each unique word is represented by a vector distributed across a set of latent dimensions extracted from a term-co-occurrence matrix (implicitly or explicitly). The resulting matrix—or in some cases, matrices (e.g., Hamilton, Leskovec, and Jurafsky, 2016; Rodman, 2020; Enggaard et al., 2023)—is often between 50 and 500 dimensions.[6] Typically, each dimension is not interpretable by itself, but rather these dimensions are treated as locations in a "semantic space."

Word embeddings (hereafter just *embeddings*) can be derived from any corpora. Analysts may want to train entirely locally on their corpora of interest, capturing the idiosyncrasies of targeted texts, or train on a wider domain of texts within which their specific corpora falls. Researchers have also made several large pretrained embeddings easily available. Such pretrained embeddings can also be expanded to incorporate novel words in target corpora (e.g., Khodak et al., 2018), or "retrofit" to target corpora to incorporate novel meanings (e.g., Dingwall and Potts, 2018).

In addition to deciding whether to use pretrained, locally trained, or domain-trained, the latter two require decisions regarding the parameters of the model—for example, context window, loss function, etc. (Antoniak and Mimno, 2018; Rodman, 2020; Aceves and Evans, 2023). Therefore, a strength of pretrained embeddings are their accessibility and consistency across studies. Researchers find pretrained embeddings tend to capture the

generic associations on par with crowdsourced hand-coded ratings (Rodriguez and Spirling, 2021) while also increasing reliability. Nevertheless, we must be critical of the underlying corpora used to train embeddings (Dodge et al., 2021).

We set aside these important considerations to focus on *post-processing procedures*, which can be applied to any set of embeddings. For our demonstrations, we use the pretrained fastText English embeddings (Bojanowski et al., 2017). These embeddings are a vocabulary of 2 million terms and 300 dimensions.[7]

Interdisciplinary research shows a range of useful information can be recovered from these distributed representations (Fulda et al., 2017; Lazaridou, Marelli, and Baroni, 2017; Caliskan and Lewis, 2020; Jones et al., 2020; Rodman, 2020; Joseph and Morgan, 2020; Arseniev-Koehler and Foster, 2022; Stoltz and Taylor, 2021; Rodriguez and Spirling, 2021; Durrheim et al., 2022; van Loon et al., 2022). For example, Utsumi (2020) finds that concrete, abstract, spatial, temporal, perceptual, and emotional knowledge are accurately encoded in embeddings. But how do we recover these kinds of information from embeddings? Next, we outline such procedures in detail (for a comprehensive discussion of the theoretical foundations of embeddings, see Arseniev-Koehler, 2021).

Before we proceed, though, we define our terms formally. We present matrices with bold uppercase letters, vectors with bold lowercase letters, and scalars with lowercase italicized letters. Therefore, we have an embedding matrix, $\mathbf{W}$, of v_1 vocabulary of unique terms and d dimensions. The cells of the matrix are real-valued, and denote a location in a real-valued coordinate space. Each term is assigned a vector, $\mathbf{w}$, of d real numbers. We use an arrow above the word to distinguish the word's vector from the literal character string, for example, dog and $\overrightarrow{dog}$. Furthermore, we continue to use document-term matrices (DTMs) to represent a corpus. Therefore, $\mathbf{D}$ is matrix of n documents and v_2 vocabulary. We use v_1 and v_2 to reference the vocabulary lengths of $\mathbf{W}$ and $\mathbf{D}$, respectively, since the vocabularies between the embedding matrix and DTM need not be the same—which will usually be the case with pretrained embeddings.

Relation Induction With Word Embeddings

Once trained, the simplest use of embeddings is measuring the relationship between two terms. Semantic similarities are mapped to geometric *distances*. More specifically, similarity relates to both paradigmatic and syntagmatic relations. A paradigmatic relation is a word's ability to substitute for another in a given context (i.e., synonymy and antonymy), and a syntagmatic

relation is whether a word typically follows or precedes another in a string (i.e., literal co-occurrence). More generally, embeddings also get higher-order similarities: two words sharing similarities to a third will also be similar to each other, and so on.

Defining relations beyond (paradigmatic or syntagmatic) similarity involves post-processing the embeddings. We compare these procedures at length. But first, let's consider the diversity of these "relations."

One family of relations is often referred to as entailment—more specifically, logical or lexical entailment (Kafe, 2019; Roller and Erk, 2016; Vylomova et al., 2016). The most common example includes *is-a* or *type-of*, as in X *is a* Y, or more concretely, a wombat *is a* marsupial. These zero in on a kind of lexical entailment known as *hyponymy* where one side is subordinate to the other side. We can make an about-face and define *hypernymy* relations. This includes *such-as* and *for example*, as in Y *such as* X or a marsupial *such as* a wombat. Entailment can also be used to define *part-whole* relations (i.e., meronymy), *located-in* relations, and other hierarchical or categorical relations, such as a scientific article's methodological inclination (Nanni and Fallin, 2021).

A second family relates to material affordances or qualities. When defining affordant relations, one associates objects with what can be done with them: for example, "window" and "open" and "song" and "sing" (Fulda et al., 2017; Vylomova et al., 2016). Similarly, one can estimate the "modality" norms of a given object: for example, the extent we engage with an object using sight, taste, touch and so on (Chersoni et al., 2020). By defining a quality relation —say, size—we can arrange the terms associated with animals, for instance, along a continuum of prototypical bigness/smallness. Using similar procedures, we could arrange objects by assumed dangerousness, intelligence, temperature, speed, and so on (Grand et al., 2018). This has also been extended to more subjective judgments, such as tastiness, warmth, nutritiousness, autonomy, righteousness, legality, importance, beauty, and many more (Richie, Zou, and Bhatia, 2019; Arseniev-Koehler and Foster, 2022; Kozlowski, Taddy, and Evans, 2019).

Understanding how embeddings encode entailment, affordance, quality, and judgment information leads us to a final family of relations building on analogy. A well-known example is "man is to computer programmer as woman is to homemaker" (Bolukbasi et al., 2016b). With analogy, the entailment relation *is-a* operates behind the scenes—that is, "computer programmer" *is a* "man." Based on this, it is common to define a generic "gender" relation where words can be easily ranked along a continuous scale from "masculine" to "feminine," corresponding to broadly shared stereotypes

(Jones et al., 2020). Recently, social scientists have generalized this to a variety of stereotypes and social identities, such as obese–skinny, Caucasian–African American, rich–poor, liberal–conservative, white–non-white, and many more (Nelson, 2021; Kozlowski, Taddy, and Evans, 2019; Arseniev-Koehler and Foster, 2022; Taylor and Stoltz, 2020b; Stoltz and Taylor, 2021).

Relation Induction Procedures

Below, we outline several post-processing procedures for defining relations with embeddings, beginning with the simplest. These are outlined in Table 1. Each procedure ultimately relies on comparing real-valued vectors. Here, we use the cosine of the angle between two vectors to measure similarity.[8] Formally, the cosine of the angle between two vectors, $\mathbf{w}$ and $\mathbf{y}$, is the dot product of the two vectors divided by the product of their lengths:

$$\cos(\mathbf{w}, \mathbf{y}) = \frac{\mathbf{w} \cdot \mathbf{y}}{\|\mathbf{w}\|_2 \|\mathbf{y}\|_2}, \tag{1}$$

where $\mathbf{w} \cdot \mathbf{y}$ is the dot product of the two vectors (the sum of the products of two vectors):

$$\mathbf{w} \cdot \mathbf{y} = \sum_{i=1}^{d} w_i y_i = w_1 y_1 + + w_d y_d, \tag{2}$$

and $\|\mathbf{w}\|_2$ is the vector length defined by the L^2-norm of the vector (i.e., square root of the dot product of the vector with itself):

$$\|\mathbf{w}\|_2 = \sqrt{\sum_{i=1}^{d} w_i^2 = w_1 w_1 + + w_d w_d}. \tag{3}$$

This is our measure of similarity, where higher values mean closer vectors.

Table 1. Procedures for Relation Induction With Word Embeddings.

Procedure	Description
Semantic centroid	Specify a concept by averaging word vectors
Semantic direction	Define juxtaposing concepts
Semantic *N*-direction	Define multi-class juxtapositions
Semantic projection	Project embeddings onto a vector
Semantic rejection	Subtract projection matrix from embeddings
Semantic region	Any *k* cluster derived from reducing embeddings

Semantic centroids. To begin, we can use a single term as an *anchor* in the embedding space. In this scenario, all words in our vocabulary are "weighted" by their distance from this single word vector—quickly building a large dictionary (e.g., Haber, 2021; Hoover et al., 2018; Garten et al., 2018). This is the most basic form of relation induction, in that we weighted each word's relation to this single word, whatever it may be. For example, Ding et al. (2016) weight all words in a corpus of social media posts by their cosine to the names of specific drugs (e.g., fentanyl, marijuana, percocet, etc.). They then use the top 20 nearest terms as a dictionary to tag posts as either discussing or not discussing a specific drug. This is analogous to asking a sample of participants to rank how related a word is to "marijuana" on a scale of −1 to 1.

A potential drawback, however, is that we might be interested in a more generic concept that is not realized precisely in a single term. For example, Voyer, Kline, and Danton (2022a) must re-center the concept of "status" to refer to non-familial and non-gender related *social standing*, and McCumber and Davis (2022) wish to specify "nature" as connotated by "natural" and "wilderness." So, how do we *specify* meanings? Standard embeddings can be "fine-tuned" by defining a new vector through some combination of existing word vectors, that is, creating a *semantic centroid* (Li, Ouyang, and Li, 2019).

To create a centroid, embedding vectors are sometimes concatenated or summed (Rossiello, Basile, and Semeraro, 2017). But the more common method (sometimes called additive) takes the average of anchor terms—for example, averaging $\overrightarrow{bank}$ and $\overrightarrow{river}$ versus $\overrightarrow{bank}$ and $\overrightarrow{money}$ (e.g., Lazaridou, Marelli, and Baroni, 2017; Grand et al., 2018).[9]

There are also several ways to average the vectors. The most common is the arithmetic mean[10] of each dimension in our vector to arrive at a semantic centroid, **c**:

$$\mathbf{c} = \frac{1}{n_a}\sum_{i=1}^{n_a}\mathbf{W}_i = \frac{\mathbf{w}_1 + \mathbf{w}_2 + \cdots + \mathbf{w}_{n_a}}{n_a}, \tag{4}$$

where n_a is the number of anchor terms, **W** is an embedding matrix, and $\mathbf{w}_{i\cdots n_a}$ are the respective vectors of **W** corresponding to the anchor terms. The result is a vector in the same dimensions as the original embeddings. We can, therefore, easily compare all the other word vectors' distances to this centroid using cosine.

As Nelson (2021: 4-5) shows, we can use centroids to represent the intersection of social identities. Using a list of terms for woman, man, Black and White, Nelson finds the arithmetic mean for each gender–race combination: black + women, black + men, white + women, and white + men.

Then, Nelson weights words associated with four domains—polity, economy, culture, and domestic—by their respective cosine to each gender–race centroids (see also Lawson et al., 2022).

Semantic directions. A potential issue with semantic centroids is that even diametrically opposed concepts will be located (relatively) close in the embedding space (Mohammad et al., 2013; Taylor and Stoltz, 2020b). For example, let's say we define a centroid for various "poverty" related vectors (e.g., $\overrightarrow{poor}$, *impoverished*). The vectors *wealth* or *rich* will likely be close to this centroid. After all, when we talk about poverty, we use words often used in discussions of affluence, such as money. This is an issue, then, only if we are interested in poverty *as opposed to* wealth. To get this juxtaposition, we define a *semantic direction*[11] pointing toward poverty and away from wealth.

Semantic directions point toward one pole of a set of meanings that are juxtaposed. They are defined using sets of opposing terms, "anchoring" both sides of the directions. Such a concept, or concept pairing, is typically understood in terms of *more-or-less*, for example, big to small, good to bad, old to young. However, directions have also been used to define lexical entailment, like hypernymy (Vylomova et al., 2016).

A widely used example is the gender spectrum, from more feminine to more masculine (Bolukbasi et al., 2016a). This idea undergirds the well-known "king is to man as queen is to woman" analogy task using the *vector offset* method introduced by Mikolov, Yih, and Zweig (2013b). First, we create a vector offset by subtracting $\overrightarrow{man}$ from $\overrightarrow{woman}$. The resulting vector defines a direction pointing toward "woman" and away from "man." Next, if we measured the cosine of "queen" and "king" with this new relation vector, $\overrightarrow{queen}$ would be closer while *king* would be further. If the cosine is zero—that is, the angle is perpendicular/orthogonal—they are equidistant to both man and woman. Therefore, this method has been used extensively to measure the extent target terms are *biased* toward man or woman (Bolukbasi et al., 2016a; Caliskan, Bryson, and Narayanan, 2017; Ornaghi, Ash, and Chen, 2019; Jones et al., 2020; Arseniev-Koehler and Foster, 2022).[12]

To increase the accuracy of this direction, we can find several gendered vectors corresponding to man on the one hand (e.g., $\overrightarrow{gentlemen}$, $\overrightarrow{boys}$) and woman on the other hand (e.g., $\overrightarrow{ladies}$, $\overrightarrow{girls}$), and then summarize the vector offset in one of three ways.

The first method, which we call "paired," involves subtracting the vectors for each juxtaposing pair, and then averaging the result of these offsets (e.g., Kozlowski, Taddy, and Evans, 2019; Taylor and Stoltz, 2020b;

Arseniev-Koehler and Foster, 2022). Therefore, using the paired method we define a semantic direction, $\mathbf{d}$, as the arithmetic mean of a set of vector offsets between a collection of juxtaposed word pairs, P:

$$\mathbf{d} = \frac{\sum_{p}^{P}(\mathbf{p_1} - \mathbf{p_2})}{|P|}, \tag{5}$$

where p is a word pair in the total set of P, $\mathbf{p_1}$ and $\mathbf{p_2}$ are the respective vectors of the two words in juxtaposed pair p, and $\mathbf{d}$ points toward $\mathbf{p_1}$ and away from $\mathbf{p_2}$.

A second method for defining a semantic direction, which we call "pooled," entails averaging all the vectors within their respective sets (Larsen et al., 2016; Arseniev-Koehler and Foster, 2022; Best and Arseniev-Koehler, 2022; van Loon et al., 2022). This creates two centroids, one for each pole of the direction. Then we offset these centroids.[13] Using the pooled method, we define a semantic direction, $\mathbf{d}$ as the arithmetic mean of a set of vectors B subtracted from the arithmetic mean of a set of vectors A:

$$\mathbf{d} = \frac{1}{|A|}(\mathbf{a}_1 + \mathbf{a}_2 + \cdots + \mathbf{a}_n) - \frac{1}{|B|}(\mathbf{b}_1 + \mathbf{b}_2 + \cdots + \mathbf{b}_n), \tag{6}$$

where $\mathbf{a}$ is a single vector corresponding to a set of A words toward which the direction will point, and $\mathbf{b}$ is a single vector in the set of B words away from which the direction will point.

A final method for defining semantic directions involves creating a set of offset vectors, as with the paired method, and then applying principal components analysis (PCA) on the resulting matrix of offset vectors (Bolukbasi et al., 2016b). The first component—explaining most of the variation among the set of offsets—is used as the direction.

More formally, let the set of vector offsets (i.e., each $[\mathbf{p_1} - \mathbf{p_2}]$) comprise the columns of a $d \times p$ matrix $\mathbf{P}$, where d is the dimensionality of the embedding matrix and p is the total number of juxtaposing word pairs. We then standardize each column of $\mathbf{P}$ to find the matrix of z-scores, $\mathbf{P}_z$, and then apply singular value decomposition to this matrix:

$$\mathbf{P}_z = \mathbf{G}\lambda\mathbf{S}^T, \tag{7}$$

where $\mathbf{G}$ is the $d \times r$ matrix of left singular vectors and $\mathbf{S}$ is the $p \times r$ matrix of right singular vectors, and, where r is the number of principal components. The lambda matrix, λ, is a diagonal matrix with singular values in decreasing order. The semantic direction, $\mathbf{d}$, is then the first singular

vector column (i.e., the first "principal component") of $\mathbf{G}$. This principal component satisfies:

$$\mathbf{\Omega d} = \lambda_d^2 \mathbf{d}, \tag{8}$$

where λ_d^2 is the eigenvalue corresponding to the semantic direction $\mathbf{d}$ (i.e., the largest eigenvalue) and $\mathbf{\Omega}$ is the $d \times d$ square matrix found with $\mathbf{P}_z \mathbf{P}_z{}^T$.

Each method is likely to give similar (if not the same) results, but there are important differences. As opposed to the paired and PCA method, the pooled method does not require pairing each word with exactly one other juxtaposing word, and could even be sets of different sizes. The PCA method, in contrast to the paired and pooled method, may be used to determine whether a single direction adequately summarizes the offsets and can even be used to define more than one direction (as we discuss later).

Regardless of the method, though, the resulting vectors are in the same dimensions as our embeddings.[14] This means we can use cosine to measure each word vector's distance from this direction. We can also easily generalize this procedure to domains beyond gender. All we need to estimate a semantic direction is a list of juxtaposing anchor terms. Just as one could pick any words in the embedding space to create a centroid, the directions one can create are only limited by the terms available in the embedding space. These anchor dictionaries can be precompiled or built by the analyst from prior theory and research (Fellbaum, 1998).

Sociologists have used directions to explore broad sociocultural patterns. For example, Kozlowski, Taddy, and Evans (2019) estimated directions for affluence and education, and Arseniev-Koehler and Foster (2022) estimate directions for health and morality. Semantic directions can also be used to define material features and affordances of objects. Grand et al. (2018) estimated a direction from $\overrightarrow{small}$ to $\overrightarrow{big}$ and show animals, for instance, are distributed intuitively along it by their prototypical size. Importantly, as they show, two vectors may be closely positioned along this semantic direction, such as $\overrightarrow{horse}$ and $\overrightarrow{tiger}$, and yet far apart in the original embeddings. Furthermore, this "size" semantic direction is agnostic to the actual categories used. We could just as easily substitute animals for foods or cities.

Semantic N-directions. A possible limitation of semantic directions is the implication that relations are roughly bipolar (Osgood, Suci, and Tannenbaum, 1957). Perhaps the meaning of gender, for example, is adequately represented as a linear spectrum, but perhaps a higher geometry better captures gendered relations in a given corpus. In this vein, work seeking to debias embeddings

has begun to focus on "multiclass" biases (Manzini et al., 2019; Schlender and Spanakis, 2020). Drawing on this literature, we can arrive at a few methods to define *semantic N-directions*.

The first, and simplest, approach involves defining centroids for each "class"—or pole in our *N*-polar structure. We then select one centroid as the reference class. Consider religion, for example, where we could define a one-dimensional "Christian" to "non-Christian" direction by selecting $\overrightarrow{christian}$ as our reference centroid and subtracting the centroids for all non-Christian centroids. This would entail lumping together "Muslim," "Jewish," "Jainist," and so on. We can repeat this process, creating a new direction, with each centroid serving as a reference class. Each term in our embeddings can then be weighted by their cosine to *n* directions within the same domain, here religion. We might call this the "pooled" offset method (cf. Garg et al., 2018).

A second method, following Bolukbasi et al. (2016b), uses PCA. This would follow the PCA procedure described in the previous section, but rather than only returning the first principal component, we could return *n* components. In this scenario, *n* may be a pre-specified number of classes, or we could use standard PCA diagnostic techniques to determine the appropriate number in terms of how well the variation in vector offsets is summarized by the fewest classes. Again, each word can be weighted by their cosine to these *n* directions.

A third method (Nanni and Fallin, 2021) involves finding centroids for each anchor set, but instead of finding the difference between a centroid and the mean of all other centroids like the pooled method, we select a single centroid as our *reference* and find each pairwise semantic direction from that reference. For instance, if centroid #1 is the reference, then we have $\mathbf{d_{1-2}}, \mathbf{d_{1-3}}, \ldots, \mathbf{d_{1-n}}$ semantic directions. Next, instead of using cosine, we model our word vectors $\mathbf{w}$ as a function of these semantic directions using a standard ordinary least squares (OLS) estimator:

$$\mathbf{w} = \sum_{j \neq 1} \beta_{1-j} \mathbf{d_{1-j}} + \varepsilon. \tag{9}$$

Note this can be understood geometrically as the dot product between $\mathbf{w}$ and $\mathbf{d_{1-j}}$, with $\mathbf{w}$ being normalized and $\mathbf{d_{1-j}}$ being centered and normalized. Which centroid is the reference is irrelevant (Nanni and Fallin, 2021): the coefficient for each *j* remains the same. Further, the coefficient for the reference centroid is also defined in this single model: $\beta_1 = \sum_j \beta_{1-j}$. Therefore, we can simplify the notation from β_{1-j} to β_j and note that β_1 is also defined.

We then normalize each β to make them comparable (Nanni and Fallin, 2021):

$$\beta'_1 = \left(\sum_j \beta_{1-j}\right) \times \|\mathbf{c_1}\|_2, \tag{10a}$$

$$\beta'_j = \beta_{1-j} \times \|\mathbf{c_j}\|_2, \tag{10b}$$

where $\mathbf{c_1}$ and $\mathbf{c_j}$ are the centroids for the first and *j*th set of words, respectively, and $\|\mathbf{c_1}\|_2$ and $\|\mathbf{c_j}\|_2$ are the L^2-norms. Then, each normalized coefficient is the similarity of $\mathbf{w}$ to the *j*th pole *controlling for similarities to all the other poles*. If *j* is not the reference centroid, then negative values indicate $\mathbf{w}$ is closer to the *j* pole; if *j is* the reference, then positive values (i.e., a positive $\sum_{i \neq j} \beta_i$) indicate $\mathbf{w}$ is closer to the *j* pole.

Semantic projections and rejections. Finding cosine involves *projecting* one vector onto a *line*—a one-dimensional "subspace." It acts like a flashlight directly behind a vector, casting its shadow onto a wall. Typically, we want the angle as a summary measure, though we could also recover the relative coordinates of this projection. Such coordinates would be in the same dimensions as the embeddings. We do so by finding the *projection matrix* for a given vector. Formally, the projection matrix, $\mathbf{\Theta}$, is:

$$\mathbf{\Theta} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{v}, \tag{11}$$

where $\mathbf{W}$ is the original embedding matrix and $\mathbf{v}$ is the vector onto which we are projecting the matrix. This vector can correspond to a single word (i.e., a $\mathbf{w}$) or any of the other relations discussed here. This projection matrix can be used for a variety of ends. For example, Yıldırım and Yıldız (2018) create a projection matrix emphasizing hypernymy relations. The projection matrix can also be used to obtain the *rejection matrix* (Roller and Erk, 2016). This is the resulting matrix when subtracting the projection matrix from the original embedding matrix.

Continuing with the example of gender, our rejection matrix neutralizes the association each word has with a semantic direction by finding word vectors orthogonal to the direction. For example, in the original embeddings, $\overrightarrow{she}$ and $\overrightarrow{mother}$ would be closer to each other than they are to $\overrightarrow{he}$ and $\overrightarrow{father}$. In the rejection matrix, $\overrightarrow{she}$ is now closer to $\overrightarrow{he}$, and $\overrightarrow{mother}$ closer to $\overrightarrow{father}$. Furthermore, the association between certain words in the original space—say "perky" and "spunky"—may be reduced when gender is rejected (Schmidt, 2015).[15]

This idea forms the backbone of attempts to "debias" a set of embeddings (Gonen and Goldberg, 2019; Manzini et al., 2019; Schlender and Spanakis, 2020). The social scientist, however, could use this gender rejection matrix to, for example, *control* for gendered associations when estimating any number of semantic domains, such as sexuality, career, or politics. In other words, we can see how much a given association in our embeddings is driven by gender. Rejection is then easily generalized beyond gender. For example, Ding et al. (2016) used this technique to distinguish between *medical* uses of drugs like fentanyl and *illicit* uses.

Semantic regions. Embeddings are, technically, a *low-dimensional* representation of otherwise high-dimensional word co-occurrences in a corpus. But embeddings still tend to be hundreds of dimensions. As a result, there is a growing literature proposing dimension-reduction techniques for embeddings. On the full embeddings, this can accurately be described as a form of topic modeling (Sia, Dalmia, and Mielke, 2020; Arseniev-Koehler et al., 2021; Arora et al., 2016a; Zhang et al., 2021), as one is learning more encompassing latent themes, which we call *semantic regions*. Of the procedures discussed, finding a region is the only one that does not require an anchor dictionary.

Semantic regions can be used to measure *indirect* biases. For example, Gonen and Goldberg (2019) argued that standard debiasing techniques based on the gender semantic direction may eliminate the gendered biases between this direction and "gender-neutral" target words. However, gendered biases remain in the associations among these target words. For example, while $\overrightarrow{petite}$ and $\overrightarrow{delicate}$ may not be more or less associated with $\overrightarrow{woman}$ as opposed to $\overrightarrow{man}$ after debiasing, these vectors tend to be associated with other stereotypically feminine terms. To define this *indirect* gender relation, Gonen and Goldberg find the 1,000 most gender-biased vectors in the original embeddings, 500 for both sides of the gender direction. They then cluster these 1,000 vectors into two groups using k-means. Even after debiasing using the technique outlined by Bolukbasi et al. (2016b), the vectors for these two indirect bias clusters accurately recover gender bias (see also Du and Joseph, 2020).

Anchors, Test Sets, and Hand-Weighted Dictionaries

Following prior work, we can use a word classification task to validate how well a defined relation captures intended meanings (Arseniev-Koehler and Foster, 2022; Best and Arseniev-Koehler, 2022). We want to measure how

accurately we classify words that we expect to be very close to a given relation. But, which words do we use in our test set? It is here where hand-weighted (expert or crowdsourced) dictionaries can be indispensable resources (Ornaghi, Ash, and Chen, 2019; Best and Arseniev-Koehler, 2022).

Using the example of sensorimotor norms, we define semantic centroids for each modality (auditory, gustatory, haptic, olfactory, and visual). Given that all modalities are likely to be semantically related to each other (e.g., smell would be close to taste), we then define semantic *N*-directions for each modality using the pooled offset method.

We then build a test set with the Lancaster Sensorimotor Norms dictionaries (see the Appendix in the supplemental material for details). As words may be rated high on multiple modalities, we use "modality exclusivity" scores (Lynott et al., 2020: 1279) to limit the candidates for our test sets.[16] We then selected the 80 terms highest on each modality, removing words included in the modality's respective anchors, as our test set. For classification tasks using cosine similarity, it is common to use any cosine above 0 as a positive classification. Since we are using words with the highest crowdsourced ratings, we set the bar considerably higher to be more conservative. The threshold we select for accurately classifying test words is a cosine similarity to the *N*-direction falling in the top 30 percent of all 2 million words in the embeddings.[17]

The accuracy—that is, the percent of terms in the Lancaster Sensorimotor dictionaries that have cosine similarities to the respective semantic *N*-direction above the threshold—are reported in Table 2. As the results show, high similarity to a modality's semantic *N*-direction is highly predictive of a term's modality rating in the Lancaster dictionary. This is the case across all six of the modalities, with the lowest accuracy being the haptic dimension (84.5 percent). Relation induction methods such as semantic *N*-directions, therefore, appear to do a good job of "tapping into" the same sensorimotor concepts captured by the crowdsourced Lancaster dictionaries.[18]

Table 2. Anchor Terms for Sensorimotor Semantic *N*-Directions.

Modality	Anchor Terms	Accuracy
Auditory	Auditory, hear, ear, and audible	100%
Gustatory	Austatory, taste, tongue, and mouth	96.9%
Haptic	Haptic, touch, hands, tactile, grip, and touching	84.5%
Olfactory	Olfactory, smell, nose, and aroma	100%
Visual	Optic, color, colorful, show, see, and vision	87%

Relation Induction at the Document-Level

Each relation induction procedure discussed allows us to measure each word's association with a relation. We can also, of course, measure the association between the relations themselves. For example, Jones et al. (2020) measured the association between masculine and feminine centroids, on the one hand, and career, science, arts, and family centroids, on the other hand (see also Leschke and Schwemmer, 2019). Kozlowski, Taddy, and Evans (2019) measured the association between various semantic directions for social class, such as cultivation, education, and affluence. Similarly, Arseniev-Koehler et al. (2021) measured associations between semantic regions and a gender direction. As embeddings are, in essence, summaries of the full training corpus, these studies explore corpus-level semantic patterns.

Analysts, however, are often interested in aggregates of text above the level of words or phrases but below the corpus-level (hereafter just *documents*). How do we use embeddings to measure the extent documents "engage" with a given relation? We describe two techniques below.

Document Centroids

First, we could use semantic centroids to represent text aggregates (Mihaylov and Nakov, 2019; Kusner et al., 2015; Brokos, Malakasiotis, and Androutsopoulos, 2016; Arora, Liang, and Ma, 2016b). For example, Berry and Taylor (2017) used this method to create a single vector representation for each comment in a corpus of Facebook Pages. Similarly, Lix et al. used this technique to represent messages in Slack, which "tends to be brief and conversational, with individual posts often comprising just a few words" (Lix et al., 2020: 10).

To get the document centroids, we can use basic matrix multiplication. First, following the notation laid out earlier, we create a document-term matrix, $\mathbf{D}$, where each row is a document, n, and each column a unique word in the vocabulary, v_2. $\mathbf{D}$ is then normalized to obtain relative frequencies (typically counts divided by the L^1-norm).[19] We'll then limit our vocabulary to those terms that intersect with the vocabulary in $\mathbf{W}$. We'll call this normalized and trimmed DTM $\mathbf{D}_\alpha$. We also have our embedding matrix, $\mathbf{W}$, of v_1 vocabulary and d dimensions. Like $\mathbf{D}_\alpha$, we trim the vocabulary of the embedding matrix so that only the terms that intersect with the vocabulary of the DTM are retained. We'll call this trimmed embedding matrix $\mathbf{W}_\alpha$. The vocabularies—that is, the columns of $\mathbf{D}_\alpha$ and the rows of $\mathbf{W}_\alpha$—are now the same and in the same order

(which we can denote simply as $\mathbf{v}$). We then find the documents' centroid vectors $\mathbf{C}$ with:

$$\mathbf{C} = \mathbf{D}_\alpha \mathbf{W}_\alpha. \tag{12}$$

The end product is a matrix $\mathbf{C}$ of n documents by d dimensions.

Using centroid representations, one could measure each document's association with relations defined with any of the procedures discussed. For example, using a liberal-conservative semantic direction, tweets could be arranged by their relative association with either pole of the direction to obtain a continuous classifier for political lean. The strength of this approach is that it is intuitive and computationally efficient, but tends to be less suited for longer sequences.[20]

Mover's Distance

To address longer documents and variation in length, we can use techniques centered on earth mover's distance (EMD) which compares two probability distributions (a.k.a. Wasserstein, Kantorovich, or Mallows' Distance) (Kantorovich, 1960; Rubner, Tomasi, and Guibas, 1998). EMD can be used to compare overall similarity between two documents—that is, word mover's distance (Kusner et al., 2015), and measure document-level similarity to induced relations—that is, concept mover's distance (Stoltz and Taylor, 2019; Taylor and Stoltz, 2020b).[21]

Martin-Caughey (2021), for example, uses this approach to measure the similarity between the detailed occupational descriptions provided by the respondents to the General Social Survey and their respective job titles. Zhang et al. (2021) used it to find the optimal k semantic regions to summarize corpus phrases (see also Gülle et al., 2020). Taylor and Stoltz (2020a) created Likert-type measures of conceptual engagement with it, Carbone and Mijs (2022) measured how much 3,660 most popular songs across 23 European countries discuss inequality, Batzdorfer et al. (2021) identified the narrative motifs used by conspiracy theorists on Twitter (see also Akram, 2020), Voyer, Kline, and Danton (2022a) measured the change in class distinction-making in nearly a century of etiquette manuals, and McCumber and Davis (2022) mapped the meanings of "nature" in elite travel journalism.

We can conceive of any document as a cloud of points in the embedding space (Stoltz and Taylor, 2021: 7), where each point is a word. The algorithm adjusts for differences in document length so that we can compare the "cost" of moving a cloud of, say, tens of thousands of points to a cloud containing a

single point. Cost is a function of weight (relative frequency of a word in a document) and distance (cosine similarity between words vectors) (Kusner et al., 2015).

The inputs to EMD are a relative frequency DTM and an embedding matrix. Only the shared vocabulary between the two matrices is used, and any new semantic relations defined are appended to the embeddings as "pseudo" word vectors and appended to the DTM as "pseudo" documents (Stoltz and Taylor, 2019). If documents are semantically similar to a given relation, however defined, the cost of moving all its comprising words is lower than if the document is very dissimilar. This addresses both the issue of repeating words by weighting with relative frequency and addresses length variation by constraining all documents to sum to one. More importantly, by using the embedding space to measure the similarity between words, it overcomes the assumption of term independence in token-counting methods while typically avoiding loss of information from dropping OOV terms from the DTM.

Formally (Atasu et al. 2017: 2, see also Rubner, Tomasi, and Guibas 1998), two documents, $\mathbf{n}_1$ and $\mathbf{n}_2$, are represented as vectors of relative frequencies; this is equivalent to a probability distribution over the vocabulary v for each document n_i, and, therefore, the sum of their vectors equals 1. EMD finds the proportional "weight" and "cost" of terms p in n_1 flowing to terms q in n_2:

$$\mathrm{EMD}_{\mathrm{n_1,n_2}} = \min\left(\sum_{\mathrm{p,q}} \mathrm{F_{p,q}} \mathrm{cos_{dist}}(\mathbf{p}, \mathbf{q})\right), \tag{13}$$

where

$$\sum_{q} F_{p,q} = n_{1(p)}, \tag{14a}$$

$$\sum_{p} F_{p,q} = n_{2(q)}. \tag{14b}$$

The row vectors of the resulting flow matrix, $\mathbf{F}$, between n_1 and n_2 must sum to the relative frequency of the terms p in n_1 (i.e., $\sum_q F_{pq} = n_{1(p)}$) (see Kusner et al., 2015: 3). Similarly, the column vectors of $\mathbf{F}$ must also sum to the relative frequency of the terms q in n_2 (i.e., $\sum_p F_{pq} = n_{2(q)}$).

Solving the full model is computationally demanding. Therefore, applications of this technique typically use the "relaxed word mover's distance" (Kusner et al., 2015), and more recent advances (Atasu et al., 2017).

These efficient approximations result in a single (scalar) output for each document by relation, or document-by-document, similarity.[22]

Document-Level Relation Induction and the OOV and Distributional Problems

Using the relation induction techniques we discuss is more efficient than building a large weighted dictionary by hand. Using only a few anchor terms, we can define (and redefine) any number of relations. Furthermore, relation induction techniques mitigate two problems: the OOV problem by allowing words that are relevant to the concept but unlikely to be included in a precompiled dictionary to influence concept engagement, and (2) the risk of a relation engagement measure being unduly influenced by document length when aggregating to the document level. We illustrate how document-level relations address these problems here.

Recall that our vocabulary for any measure is the overlap between our corpus vocabulary and either the precompiled dictionary or the word embeddings. Embeddings trained on the corpus will have the same vocabulary, and large pretrained models will likely have a larger vocabulary and thus probably include more of the corpus vocabulary than a precompiled dictionary.

The DoCA DTM has a vocabulary of roughly 66,000 terms. The intersection of this vocabulary with the fastText word embeddings is roughly 50,000 (75 percent) and roughly 15,000 with Lancaster dictionaries (23 percent). Lemmatizing the DTM to maximize this overlap reduces the DTM to 55,000 terms, but increases the overlap (28 percent).

We can assess whether the dictionaries are missing highly relevant words. We selected the highest-rated words along each modality from the Lancaster dictionaries, and then find the cosine similarity of these words to all the words in our 50,000 word vectors present in both the embeddings and the corpus. Are there terms that are highly similar to these focal modalities,[23] but that are absent from the Lancaster dictionary? Table 3 shows both the percentage of the top 200 words that are not in the Lancaster dictionary and the top 10 words that are not in the dictionary. Overall, around 20 percent to 40 percent of these words are absent. In addition, many of these absent words seem to be highly relevant for describing the modality in question.

Building on the capacity of word embeddings to incorporate more of the vocabulary, along with cosine offering a more fine-grained, continuous measure than Likert-type ordinal rankings, relation induction techniques should also allow us to effectively unmoor the relationship between document length and concept engagement. To test this, we conducted a

Table 3. Embeddings and the Out-of-Vocabulary (OOV) Problem.

Modality	Prop. Top 200 Words Not in Dictionary	Top Ten Non-Dictionary Words
Auditory	43.5%	decibels, louder, loudspeakers, deafening, muffled, tonal, loudest, noises, cacophony, and tones
Gustatory	43.5%	alfredo, pizzas, salads, vegetables, dishes, cooked, feta, italian, tomatoes, and potatoes
Haptic	29%	acknowledgment, gestures, goodbye, kissing, pleasantries, greetings, grudging, smiles, tacit, and kisses
Olfactory	47%	incenses, candles, joss, scented, perfumes, copal, scents, spices, altars, and attar
Visual	21%	seeing, seen, happens, noticed, don, surprised, wondering, could, sees, and wondered

"random corpus" (RANCOR) analysis involving simulating RANCORs based on empirical parameters of existing corpora. In addition to the DoCA corpus, we include two additional corpora—a collection of 6,027 articles from the *American Sociological Review* (*ASR*) and just over 2.8 million speech chunks from the *Congressional Record* (*CR*). We also use two other weighted dictionaries in addition to the modality dictionaries: a concreteness dictionary (Brysbaert, Warriner, and Kuperman, 2014) and a sentiment dictionary (Rinker, 2022). (All corpora and dictionaries are described in the Appendix in the supplemental material)

The RANCOR procedure is outlined in detail in the Appendix in the supplemental material. In short, a total of 3,000 random DTMs were generated: 1,000 each using vocabularies and term probabilities from the three real corpora, and each DTM consisting of 1,000 synthetic documents. The term frequencies in the DTMs per each of the three sets were sampled such that each DTM had means, minimums, and maximums randomly sampled from a parent normal distribution with a mean, minimum, and maximum equal to the observed corpus. The only parameter of the parent normal distribution allowed to vary was the standard deviation of the document lengths, which could range from four times smaller than the standard deviation of the real corpus and four times larger than the standard deviation of the real corpus.

Then, for each random DTM, a series of seven CMD scores were generated: five semantic N-directions for the sensorimotor dictionaries, a

concreteness semantic direction (with concreteness and abstractness poles), and a polarity semantic directions (with negative and positive poles). Finally, for each of the 3,000 random DTMs, the standard deviations of the document lengths and each CMD score were generated. The goal, then, was to see if variance in document length is associated with variance in CMD scores.

Figure 2 visualizes the results. We found that, all other parameters (mostly) equal, variance in CMD-based modality scores is not meaningfully associated with variance in document length for the three corpora (as illustrated by the relatively flat smoothed trend line).

Finally, returning to just the DoCA and sensorimotor modalities, Table 4 shows select quotes from the articles with the highest CMD scores for each modality. The excerpts demonstrate considerable face validity for the measure. However, these same articles are ranked quite low using RF-based weighted dictionary scores (i.e., the product of the article-specific term frequencies with their respective dictionary weights, summed, and then divided by the article length). Indeed, the Spearman's ρ (last column) between the two measures is relatively low overall.

Discussion

In this article, we show (1) how word embeddings can be used to extract a range of gradational relations in ways that overcome some of the limitations of traditional weighted dictionary methods, and (2) how weighted dictionaries can be used to build and validate these relations. We provide a comprehensive overview of relation induction techniques that offer gradational measures of semantic similarity, hierarchy, entailment, and stereotype at the document- and author-level.

When inducing a relation from an embedding space, analysts typically rely on previous literature, their own expertise, and a thesaurus to build anchor dictionaries. Alongside this, analysts could construct a small dictionary of hand-coded words and test if they are among the relation's nearest neighbors (e.g., Arseniev-Koehler and Foster, 2022: 1533-539). Others have used research assistants or crowdsourced workers to build test dictionaries (Gennaro and Ash, 2021; Kozlowski, Taddy, and Evans, 2019; Durrheim et al., 2022). We can also use precompiled hand-weighted dictionaries, such as the Lancaster dictionaries used here, to build test sets. Ornaghi, Ash, and Chen (2019), for example, use a test set drawn from the Linguistic Inquiry and Word Count (LIWC) Dictionaries (Tausczik and Pennebaker, 2010) to validate their gender and career-family semantic directions.

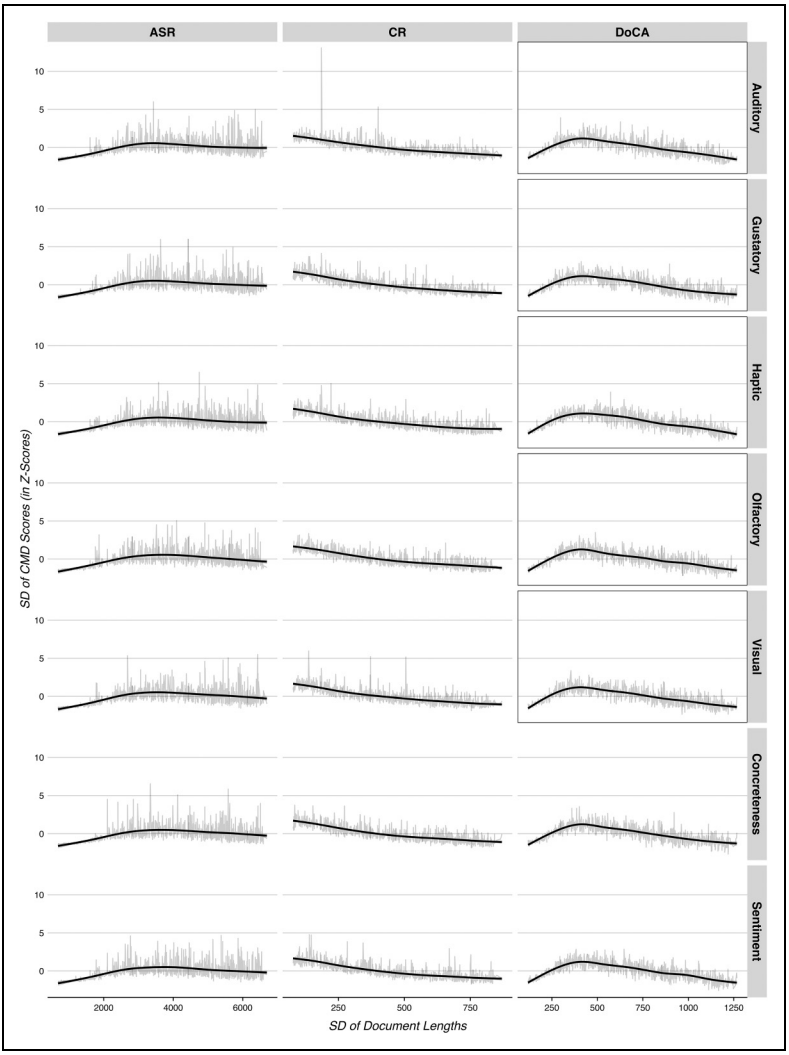

Figure 2. Concept Mover's Distance random corpus (CMD RANCOR) analysis.

At aggregate levels, the task involves document classification by checking whether key document-level covariates track engagement as expected. For example, where political speeches fall on the liberal-conservative semantic direction—either using centroid representations of these text aggregates or

Table 4. Articles With Highest CMD Scores, per Modality.

Modality	Quote	RF Rank	Spearman's ρ
Auditory	"Deaf Actress's Use of Speech Proves Divisive Among Peers"[1]	2,980	−0.09***
Gustatory	"Circuit Court [was] asked to enjoin 12 large companies from polluting Lake Michigan by dumping industrial wastes…where Chicago obtains much of its water supply"[2]	11,731	−0.02**
Haptic	"G.E. Resists War Protest; Honeywell Bars Arms Halt"[3]	15,418	0.09***
Olfactory	"Say goodbye to crayons that smelled good enough to eat."[4]	12,902	−0.07***
Visual	"A Russian health exhibition opened today in the Music Hall while about a score of pickets paraded outside."[5]	13,733	0.25***

Note: Each quote comes from the article with the highest CMD score on that modality, and rank of that same article with the RF-weighted dictionary score (out of 17,493 articles). Spearman's ρ is the rank correlation between the CMD and RF ranks for all articles on each modality. CMD = Concept Mover's Distance; RF = relative frequency.
[1]Wilson (1988); [2]*New York Times* (1967); [3]*New York Times* (1972); [4]Lawson (1995); [5]*New York Times* (1965).
**$p < .01$; ***$p < .001$ (two-tailed tests).

concept mover's distance—can be compared against ideology scores based on roll-call votes that situate individual legislators in ideological space (Gennaro and Ash, 2021; Rheault and Cochrane, 2019). Similarly, analysts can read a subset of documents to check the performance and face validity of automated classifications, in addition to guiding close readings (Rodman, 2020; Carbone and Mijs, 2022; Batzdorfer et al., 2021; Voyer et al., 2022b; McCumber and Davis, 2022).

As robustness checks for both word and aggregate levels, we should consider sensitivity to the specific method used to define the relation against some of the alternatives detailed here. Researchers can turn to formal metrics to capture the quality of derived relations, for instance, as detailed by Boutyline and Johnston (2023). This also implicates all the many preprocessing steps that text analysts may take in preparing texts for classification. As there are many steps to consider, we anticipate further advances in standardizing word embedding procedures.

Finally, it is important to reiterate that the limitations of dictionary methods discussed here applies to dictionaries used in specific ways

(as opposed to all dictionaries and count methods), and analysts should be mindful of the tradeoffs to transparency and simplicity when turning to embeddings methods, or similar. Relation induction offers a technical complement to dictionary approaches as an unsupervised, deductive method. Research using word embeddings is relatively nascent in the social sciences with recent examples coming from sociology (Jones et al., 2020; Nelson, 2021; McCumber and Davis, 2022), political science (Rheault and Cochrane, 2019; Rodman, 2020; Rodriguez, Spirling, and Stewart 2023), economics (Gennaro and Ash, 2021; Leschke and Schwemmer, 2019; Ornaghi, Ash, and Chen, 2019), and psychology (Garten et al., 2018; Hoover et al., 2018). As social scientists turn to large corpora such as social media data, decades of newspaper articles, newly digitized archives, congressional records, movie subtitles, or Wikipedia entries to test theories about the social world, their toolbox must adapt. Word embeddings, and relation induction techniques discussed in this article, are a fitting addition to analysts' toolbox.

Author's Note

We thank the anonymous reviewers for substantially improving this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

ORCID iDs

Dustin S. Stoltz https://orcid.org/0000-0002-4774-0765
Marshall A. Taylor https://orcid.org/0000-0002-7440-0723
Jennifer S. K. Dudley https://orcid.org/0000-0001-8413-6201

Data Availability Statement

The Document-Term Matrix for the Dynamics of Collective Action Corpus is available on Zenodo (Stoltz, Taylor, Dudley, 2023a) along with associated article metadata and variables from the original Dynamics of Collective Action dataset (https://web.stanford.edu/group/collectiveaction/). Code and data to reproduce the analyses in

this article are available at GitLab (https://gitlab.com/mtaylor_soc/toolkit) and as a permanent record on Zenodo (Stoltz, Taylor, and Dudley, 2023b). The analysis also relies on a few specialized R packages (Selivanov, Bickel, and Wang, 2020; Rinker, 2022; Stoltz and Taylor 2022).

Supplemental Material

Supplemental material and Appendix for this article are available online.

Notes

1. It is unfortunate that *tf-idf* is not called *rf-idf*, since the first term is, in fact, the term's relative frequency (*rf*) and not the raw count, commonly called term frequency (*tf*).
2. More specifically, the *set* of frequencies, since these are binned to aid in visualization.
3. Ranks are after terms in the 2014 Snowball stop list are omitted. We chose these terms given their substantive significance to the corpus as a protest event dataset. The other top terms show this same distributional tendency.
4. Correlations can tell us the degree a measure contributes unique information. If the "normed" frequency of a term and its "binary" absence or presence are highly correlated, we should be skeptical of *rf* and *tf-idf*, or other norming schemes, as measuring magnitude. To test this, we obtain correlations for the more than 66000 terms in the DoCA DTM between binary, on the one hand, and *rf* and *tf-idf* weightings, on the other hand, over 85 percent are above 0.70 for both *rf* and *tf-idf*, the conventional "high" correlation threshold. Thus, common norming schemes do not contribute much more than a binary measure.
5. Here, the documentation is thin, but incorporating frequency information is commonplace. In the context of sentiment analysis on movie reviews from the Internet Movie Database (IMDB), one early study did find better results with discarding frequency information (Pang, Lee, and Vaithyanathan, 2002).
6. Accuracy on certain tasks tend to improve up to 300 dimensions (see Pennington, Socher, and Manning, 2014; Rodriguez and Spirling, 2021). It may be that distance provides diminishing information as dimensions increase, see Francois, Wertz, and Verleysen (2007); Couillet et al. (2020). See also Landauer and Dumais (1997: 220-221) for this same point in the context of latent semantic analysis.
7. The training corpus is the Common Crawl from May 2017, which includes all web pages not blocked by a robot.txt protocol, totaling 630 billion tokens.
8. Cosine is the most common measure of similarity in embeddings research. This metric is intuitive and, when centered, identical to Pearson's correlation (van Dongen and Enright, 2012). It is also a holdover from early vector space

models in information retrieval. Euclidean distance is also common (Korenius, Laurikkala, and Juhola, 2007; Azarpanah and Farhadloo, 2021). In practice, we find they produce similar results, with most embedding between 100 and 400 dimensions trained with common procedures. The best metric likely depends upon the dimension of the embeddings and the clustering of vectors in the space (cf. Aggarwal, Hinneburg, and Keim, 2001; Sidorov et al., 2014; Zhelezniak et al., 2019).

9. When concatenating, vectors are *appended*. For example, concatenating two vectors of d dimensions results in a vector of $d \times 2$ dimensions. In addition to specifying concepts, centroids are also used for disambiguating word senses, creating bigrams vectors out of unigrams, representing sentences, and creating vectors for words that may not be in the embeddings (Khodak et al., 2018; Rodrigues, Spriling, and Stewart, 2023; Iacobacci, Pilehvar, and Navigli, 2016; Wieting er al., 2016; Bojanowski et al., 2017). When using context terms from a corpus (e.g., Arora, Liang, and Ma, 2016b; Khodak et al., 2018), context words are sometimes weighted by their distance from the focal word or their overall frequencies. Although averaging is common, performance likely varies by the task.
10. In the literature, we also see the geometric mean and the harmonic mean used (Chiang, Camacho-Collados, and Pardos, 2020).
11. This is sometimes referred to as a "dimension" from the more specific "one-dimensional subspace." We prefer direction, however, to evade the many meanings of "dimension" and because it describes the underlying spatial behavior. Additionally, Boutyline and Johnston (2023) refer to this as a "semantic axis" which seems an equally satisfactory label.
12. Here we define the gender direction first, then measure the cosine between terms and direction. In the literature on bias, it is common to start by measuring target terms' cosine with each gender centroid separately and then subtract the two to get target terms' relative bias.
13. Nanni and Fallin (2021) refer to this approach as the dimension creation component of "simple semantic-dimension analysis" in their methodological appendix.
14. We can even define directions across embedding spaces. For instance, Enggaard et al. (2023) use the vector offset method to define a direction between the same term used in two *separate* embeddings to contrast how two corpora talk about similar terms.
15. Generalizing the idea of first-order and second-order similarities, Artetxe et al. (2018) produce embeddings that emphasize n^{th} order similarities. As the second-order similarity matrix would be $(\mathbf{W}\mathbf{W}^T)^2 = \mathbf{W}\mathbf{W}^T\mathbf{W}\mathbf{W}^T$, we can find $(\mathbf{W}\mathbf{W}^T)^n$, for any embedding matrix $\mathbf{W}$.

16. The mean exclusivity was 0.23, and the maximum 0.63. We included words at or above 0.3.
17. In practice, the analyst could increase the accuracy by continuing to tweak the anchor terms used to create the directions, potentially using some combination of metrics of fit (e.g., Boutyline and Johnston, 2023)
18. Van Loon and colleagues (2022, see also Valentini et al., 2023) show that semantic relations (specifically, embedding association tests) can exhibit biases when the words that define one pole of a relation are positioned together with frequent words and the words for the other pole are positioned with rare words in the embedding space. To assess if these semantic *N*-directions exhibit a similar bias with term frequencies, we subset the fastText embeddings to only the columns in the DoCA DTM and then computed the Pearson's correlation between each term's term frequency and its cosine similarity to each modality-specific directions. The correlations were all quite small: no correlation had an absolute value above 0.04.
19. Alternatively, we may want to down-weight frequent words. Therefore, we could use the *inverse* of word frequency (Boutyline, Cornell, and Arseniev-Koehler, 2021; Arora, Liang, and Ma, 2016b; Karipbayeva, Sorokina, and Assylbekov, 2019).
20. Averaging or additive approaches tend to amplify the components of common dimensions at the expense of less common dimensions, which can be somewhat improved by removing very frequent words (cf. Khodak et al., 2018; Mu, Bhat, and Viswanath, 2017; Rodriguez, Spirling, and Stewart 2023).
21. One additional method, called doc2vec (Le and Mikolov, 2014), involves training embeddings at the document-level. The procedure is roughly similar to document centroids in that each document is assigned a single vector representation, see, for example, Pardo-Guerra and Pahwa (2022); Haber, Haveman, and Hong (2021); Aceves and Evans (2023).
22. These approximations typically do not retain the word-level associations which may themselves be useful (e.g., Brunila and LaViolette, 2021).
23. The focal words are as follows: "loudness" (auditory), "pasta" (gustatory), "handshake" (haptic), "incense" (olfactory), and "see" (visual). Each dictionary contained multiple words with the highest possible weighting (5), except for haptic (which had a maximum weight of 4.94). The terms selected here as focal words were simply the first top-weighted word when sorted alphabetically. The one exception to this was the visual focal word—"see"—which we chose because it seemed to be a more fair exemplar of the modality than the word that would've been chosen had we gone with top word when sorted alphabetically ("pretty").

24. We could not find the exact number. The protest event is the unit of analysis in the DoCA, not the article. Multiple articles are sometimes used to describe a single event (and one article may reference multiple events), but the codes across these articles are pooled together into a single event row. However, since we matched re-collected article names to the listed article names (with the variable called "title") in the DoCA, we only return a single article per event—that is, the article with the closest string match, using the DoCA article name as the query string.
25. Analysts tend to remove words that seem uninformative. These words are placed in a "stoplist" or "negative dictionary." How these words are selected varies. Analysts could remove the most frequent few words in their own corpus, but analysts often use precompiled stoplists. There are numerous stoplists, and they vary widely, often ranging in size from around fifty to over a thousand words. The inclusion criteria are also usually poorly documented, and many have errors (Nothman, Qin, and Yurchak, 2018). The 2014 Snowball list is by far the most common.

References

Aceves, Pedro and James A. Evans. 2023. "Mobilizing Conceptual Spaces: How Word Embedding Models Can Inform Measurement and Theory Within Organization Science." *Organization Science*: 1-27.

AkramAl-Turk. 2020. *The Rise of Performance-Based Accountability in Education in the United States: 1965-1994*. PhD thesis, University of North Carolina at Chapel Hill.

AntoniakMaria and David Mimno. 2018. "Evaluating the Stability of Embedding-Based Word Similarities." *Transactions of the Association for Computational Linguistics* 6:107-19.

AroraSanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016a. "A Latent Variable Model Approach to PMI-Based Word Embeddings." *Transactions of the Association for Computational Linguistics* 4:385-99.

AroraSanjeev, Yingyu Liang, and Tengyu Ma. 2016b. "A Simple but Tough-to-Beat Baseline for Sentence Embeddings." *5th International Conference on Learning Representations*.

Arseniev-KoehlerAlina. 2021. "Theoretical Foundations and Limits of Word Embeddings: What Types of Meaning Can They Capture?" *arXiv* 2107.10413.

Arseniev-KoehlerAlina, Susan D. Cochran, Vickie M. Mays, Kai-Wei Chang, and Jacob Gates Foster. 2021. "Integrating Topic Modeling and Word Embedding to Characterize Violent Deaths." *arXiv* 2106.14365.

Arseniev-KoehlerAlina and Jacob G. Foster. 2022. "Machine Learning As a Model for Cultural Learning: Teaching An Algorithm What it Means to be Fat." *Sociological Methods & Research* 51:1484-539.

ArtetxeMikel, Gorka Labaka, Iñigo Lopez-Gazpio, and Eneko Agirre. 2018. "Uncovering Divergent Linguistic Information in Word Embeddings With Lessons for Intrinsic and Extrinsic Evaluation." *arXiv 1809.02094*.

AslanidisParis. 2018. "Measuring Populist Discourse With Semantic Text Analysis: An Application on Grassroots Populist Mobilization." *Quality & Quantity* 52:1241-63.

AtasuKubilay, Thomas Parnell, Celestine Dünner, Manolis Sifalakis, Haralampos Pozidis, Vasileios Vasileiadis, Michail Vlachos, Cesar Berrospi, and Abdel Labbi. 2017. "Linear-Complexity Relaxed Word Mover's Distance With GPU Acceleration." Pp. 889-96 in *2017 IEEE International Conference on Big Data*, IEEE.

Baayen, Harald R.. 2002. *Word Frequency Distributions*. Berlin, Germany. Springer.

Batzdorfer, Veronika, Holger Steinmetz, Marco Biella, and Meysam Alizadeh. 2021. "Conspiracy Theories on Twitter: Emerging Motifs and Temporal Dynamics During the COVID-19 Pandemic." *International Journal of Data Science and Analytics* 13:315-333.

BerryGeorge and Sean J. Taylor. 2017. "Discussion Quality Diffuses in the Digital Public Square." Pp. 1371-380 in *Proceedings of the 26th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee.

Best, Rachel K. and Alina Arseniev-Koehler. 2022. "Stigma's Uneven Decline." *SocArXiv*. 10.31235/osf.io/7nm9x

Bhatt, Anjali M., Amir Goldberg, and Sameer B. Srivastava. 2021. "A Language-Based Method for Assessing Symbolic Boundary Maintenance Between Social Groups." *Sociological Methods & Research* 51(4):1681-1720.

BojanowskiPiotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. "Enriching Word Vectors With Subword Information." *Transactions of the Association for Computational Linguistics* 5:135-46.

BolukbasiTolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016a. "Quantifying and Reducing Stereotypes in Word Embeddings." *arXiv* 1606.06121.

BolukbasiTolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016b. "Man is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." Pp. 4349-357 in *Advances in Neural Information Processing Systems 29*.

BouraouiZied, Shoaib Jameel, and Steven Schockaert. 2018. "Relation Induction in Word Embeddings Revisited." Pp. 1627-637 in *Proceedings of the 27th International Conference on Computational Linguistics*.

Boutyline, Andrei, Devin Cornell, and Alina Arseniev-Koehler. 2021. "All Roads Lead to Polenta." *Sociological Forum* 36(S1):1419-1445.

BoutylineAndrei and Ethan Johnston. 2023. "Forging Better Axes.".

BrokosGeorgios-Ioannis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2016. "Using Centroids of Word Embeddings and Word Mover's Distance for Biomedical Document Retrieval in Question Answering." *arXiv* 1608.03905.

BrunilaMikael and Jack LaViolette. 2021. "WMDecompose: A Framework for Leveraging the Interpretable Properties of Word Movers Distance in Sociocultural Analysis." *arXiv 2110.07330*.

BrysbaertMarc, Amy Beth Warriner, and Victor Kuperman. 2014. "Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas." *Behavior Research Methods* 46:904-11.

CaliskanAylin, Joanna J. Bryson, and Arvind Narayanan. 2017. "Semantics Derived Automatically From Language Corpora Contain Human-Like Biases." *Science (New York, NY)* 356:183-6.

CaliskanAylin and Molly Lewis. 2020. "Social Biases in Word Embeddings and Their Relation to Human Cognition." *PsyArXiv* d84kg.

Carbone, Luca and Jonathan Mijs. 2022. "Sounds Like Meritocracy to My Ears: Exploring the Link Between Inequality in Popular Music and Personal Culture." *Information, Communication and Society* 25(5):707-725. 10.1080/1369118X.2021.2020870

Charu C.Aggarwal, Alexander Hinneburg, and Daniel A. Keim. 2001. "On the Surprising Behavior of Distance Metrics in High Dimensional Space." Pp. 420-34 in *Database Theory*. Springer.

Cheng, Mengjie, Daniel Scott Smith, Xiang Ren, Hancheng Cao, Sanne Smith, and Daniel A. McFarland. 2023. "How New Ideas Diffuse in Science." *American Sociological Review*. 00031224231166955.88(3):522-561.

ChersoniEmmanuele, Rong Xiang, Qin Lu, and Chu-Ren Huang. 2020. "Automatic Learning of Modality Exclusivity Norms With Crosslingual Word Embeddings." Pp. 32-38 in *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*.

ChiangHsiao-Yu, Jose Camacho-Collados, and Zachary Pardos. 2020. "Understanding the Source of Semantic Regularities in Word Embeddings." Pp. 119-31 in *Proceedings of the 24th Conference on Computational Natural Language Learning*.

CouilletRomain, Yagmur Gizem Cinar, Eric Gaussier, and Muhammad Imran. 2020. "Word Representations Concentrate and This Is Good News!" Pp. 325-34 in *Proceedings of the 24th Conference on Computational Natural Language Learning*.

DamienFrancois, Vincent Wertz, and Michel Verleysen. 2007. "The Concentration of Fractional Distances." *IEEE Transactions on Knowledge and Data Engineering* 19:873-86.

Deterding, Nicole M. and Mary C. Waters. 2018. "Flexible Coding of In-Depth Interviews: A Twenty-First-Century Approach." *Sociological Methods & Research.* 50(2):708-739.

DingTao, Arpita Roy, Zhiyuan Chen, Qian Zhu, and Shimei Pan. 2016. "Analyzing and Retrieving Illicit Drug-Related Posts from Social Media." Pp. 1555-560 in *2016 IEEE International Conference on Bioinformatics and Biomedicine*.

DingwallNicholas and Christopher Potts. 2018. "Mittens: An Extension of GloVe for Learning Domain-Specialized Representations.".

DodgeJesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus.".

DuYuhao and Kenneth Joseph. 2020. "MDR Cluster-Debias: A Nonlinear Word Embedding Debiasing Pipeline." Pp. 45-54 in *Social, Cultural, and Behavioral Modeling*.

Durrheim, Kevin, Maria Schuld, Martin Mafunda, and Sindisiwe Mazibuko. 2022. "Using Word Embeddings to Investigate Cultural Biases." *The British Journal of Social Psychology* 62(1):617-629.

EarlJennifer, Sarah A. Soule, and John D. McCarthy. 2003. "Protest Under Fire? Explaining the Policing of Protest." *American Sociological Review* 68:581-606.

EnggaardThyge, August Lohse, Morten Axel Pedersen, and Sune Lehmann. 2023. "Dialectograms: Machine Learning Differences Between Discursive Communities.".

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

FloresRené D.. 2017. "Do Anti-immigrant Laws Shape Public Sentiment? A Study of Arizona's SB 1070 Using Twitter Data." *American Journal of Sociology* 123:333-84.

FryeMargaret and Nina Gheihman. 2018. "Like Bees to a Flower: Attractiveness, Risk, and Collective Sexual Life in An AIDS Epidemic." *Sociological Science* 5:596-627.

FuldaNancy, Daniel Ricks, Ben Murdoch, and David Wingate. 2017. "What Can You Do With a Rock? Affordance Extraction via Word Embeddings." *arXiv prvolume arXiv:1703.03429*.

GargNikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." *Proceedings of the National Academy of Sciences of the United States of America* 115:E3635-E3644.

GartenJustin, Joe Hoover, Kate M. Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani. 2018. "Dictionaries and Distributions: Combining Expert

Knowledge and Large Scale Textual Data Content Analysis." *Behavior Research Methods* 50:344-61.

Gennaro, Gloria and Elliott Ash. 2021. "Emotion and Reason in Political Language." *The Economic Journal* 132(643):1037-1059.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. 2018. "Congressional Record for the 43rd–114th Congresses: Parsed Speeches and Phrase Counts." https://data.stanford.edu/congresstext.

GentzkowMatthew, Jesse M. Shapiro, and Matt Taddy. 2019. "Measuring Group Differences in High-dimensional Choices: Method and Application to Congressional Speech." *Econometrica: Journal of the Econometric Society* 87:1307-40.

GoldbergAmir, Sameer B. Srivastava, V. Govind Manian, William Monroe, and Christopher Potts. 2016. "Fitting in Or Standing out? The Tradeoffs of Structural and Cultural Embeddedness." *American Sociological Review* 81:1190-222.

GonenHila and Yoav Goldberg. 2019. "Lipstick on a Pig: Debiasing Methods Cover Up Systematic Gender Biases in Word Embeddings But Do Not Remove Them." *arXiv* 1903.03862.

GrandGabriel, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2018. "Semantic Projection: Recovering Human Knowledge of Multiple, Distinct Object Features From Word Embeddings." *arXiv prvolume arXiv:1802.01241*.

GülleKim Julian, Nicholas Ford, Patrick Ebel, Florian Brokhausen, and Andreas Vogelsang. 2020. "Topic Modeling on User Stories Using Word Mover's Distance." Pp. 52-60 in *2020 IEEE Seventh International Workshop on AIRE*.

Haber, Jaren. 2021. "Sorting Schools: A Computational Analysis of Charter School Identities and Stratification." *Sociology of Education* 94(1):43-64.

HaberJaren, Heather Haveman, and Yoon Sung Hong. 2021. "Toward Computational Literature Reviews: Applying Expert-Built Dictionaries for Automated Analysis of Complex Texts.".

HamiltonWilliam L., Jure Leskovec, and Dan Jurafsky. 2016. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." Pp. 1489-501 in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics.

Hoover, Joe, Kate Johnson, Reihane Boghrati, Jesse Graham, Morteza Dehghani, and M. Brent Donnellan. 2018. "Moral Framing and Charitable Donation." *Collabra: Psychology* 4(1):1-18. 10.1525/collabra.129

HosseinAzarpanah and Mohsen Farhadloo. 2021. "Measuring Biases of Word Embeddings: What Similarity Measures and Descriptive Statistics to Use?" Pp.

8-14 In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, Association for Computational Linguistics.

HuMinqing and Bing Liu. 2004. "Mining and Summarizing Customer Reviews." Pp. 168-77 in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

IacobacciIgnacio, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. "Embeddings for Word Sense Disambiguation: An Evaluation Study." pp. 897-907 in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers). Association for Computational Linguistics.

JiaqiMu, Suma Bhat, and Pramod Viswanath. 2017. "All-But-the-Top: Simple and Effective Postprocessing for Word Representations.".

JockersMatthew L.. 2015. *Syuzhet: Extract Sentiment and Plot Arcs from Text*.

JonesJason J., Mohammad Ruhul Amin, Jessica Kim, and Steven Skiena. 2020. "Stereotypical Gender Associations in Language Have Decreased Over Time." *Sociological Science* 7:1-35.

JosephKenneth and Jonathan H. Morgan. 2020. "When Do Word Embeddings Accurately Reflect Surveys on Our Beliefs About People?" *arXiv* 2004.12043.

KafeEric. 2019. "Fitting Semantic Relations to Word Embeddings." p. 228 in *Wordnet Conference*.

KantorovichL. V.. 1960. "Mathematical Methods of Organizing and Planning Production." *Management Science* 6:366-422.

KaripbayevaAidana, Alena Sorokina, and Zhenisbek Assylbekov. 2019. "A Critique of the Smooth Inverse Frequency Sentence Embeddings." *arXiv 1909.13494*.

KhodakMikhail, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon Stewart, and Sanjeev Arora. 2018. "A La Carte Embedding: Cheap but Effective Induction of Semantic Feature Vectors.".

KoreniusTuomo, Jorma Laurikkala, and Martti Juhola. 2007. "On Principal Component Analysis, Cosine and Euclidean Measures in Information Retrieval." *Information Sciences* 177:4893-905.

Kornai, András. 2008. *Mathematical Linguistics*. London: Springer.

KovácsBalázs, Glenn R. Carroll, and David W. Lehman. 2017. "The Perils of Proclaiming An Authentic Organizational Identity." *Sociological Science* 4:80-106.

KozlowskiAustin C., Matt Taddy, and James A. Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class Through Word Embeddings." *American Sociological Review* 84:905-49.

KusnerMatt, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. "From Word Embeddings to Document Distances." Pp. 957-66 in *International Conference on Machine Learning*.

LandauerThomas K. and Susan T. Dumais. 1997. "A Solution to Platos Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge." *Psychological Review* 104:211.

LarsenAnders Boesen Lindbo, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2016. "Autoencoding Beyond Pixels Using a Learned Similarity Metric." Volume 48, Pp. 1558-566 in *Proceedings of the 33rd International Conference on Machine Learning*.

Lawson, Carol. 1995. "After a Protest by Parents, Crayola Changes its Recipes." *The New York Times*. Nov. 15, 1995. Section C:11. https://www.nytimes.com/1995/11/15/garden/after-a-protest-by-parents-crayola-changes-its-recipes.html

LawsonM. Asher, Ashley E. Martin, Imrul Huda, and Sandra C. Matz. 2022. "Hiring Women Into Senior Leadership Positions is Associated With a Reduction in Gender Stereotypes in Organizational Language." *PNAS* 119.

LazaridouAngeliki, Marco Marelli, and Marco Baroni. 2017. "Multimodal Word Meaning Induction From Minimal Exposure to Natural Text." *Cognitive Science* 41 Suppl 4:677-705.

LeQuoc and Tomas Mikolov. 2014. "Distributed Representations of Sentences and Documents." Pp. 1188-196 in *International Conference on Machine Learning*.

Leschke, Julia C. and Carsten Schwemmer. 2019. "Media Bias Towards African-Americans Before and After the Charlottesville Rally." P. 10 in *Weizenbaum Conference*.

LiChangchun, Jihong Ouyang, and Ximing Li. 2019. "Classifying Extremely Short Texts by Exploiting Semantic Centroids in Word Mover's Distance Space." Pp. 939-49 in *The World Wide Web Conference*.

LiuBing, Minqing Hu, and Junsheng Cheng. 2005. "Opinion Observer: Analyzing and Comparing Opinions on the Web." Pp. 342-51 in *Proceedings of the 14th International Conference on WWW*.

LixKatharina, Amir Goldberg, Sameer Srivastava, and Melissa A. Valentine. 2020. "Aligning Differences: Discursive Diversity and Team Performance."

LynottDermot, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. "The Lancaster Sensorimotor Norms." *Behavior Research Methods* 52:1271-91.

ManziniThomas, Yao Chong Lim, Yulia Tsvetkov, and Alan W. Black. 2019. "Black Is to Criminal as Caucasian Is to Police: Detecting and Removing Multiclass Bias in Word Embeddings." *arXiv* 1904.04047.

Martin-Caughey, Ananda. 2021. "What's in An Occupation? Investigating Within-Occupation Variation and Gender Segregation Using Job Titles and Task Descriptions." *American Sociological Review* 86(5):960-999.

McCumber, Andrew and Adam Davis. 2022. "Elite Environmental Aesthetics: Placing Nature in a Changing Climate." *American Journal of Cultural Sociology*. doi: 10.1057/s41290-022-00179-w

MihaylovTodor and Preslav Nakov. 2019. "SemanticZ at SemEval-2016 Task 3: Ranking Relevant Answers in Community Question Answering Using Semantic Similarity Based on Fine-Tuned Word Embeddings." *arXiv* 1911.08743.

MikolovTomas, Wen-tau Yih, and Geoffrey Zweig. 2013a. "Linguistic Regularities in Continuous Space Word Representations." Pp. 746-51 in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia. Association for Computational Linguistics.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013b. "Linguistic Regularities in Continuous Space Word Representations." Pp. 746-51 in *Proceedings of the 2013 Conference of the NAACL*.

Miles, Matthew B. and Michael A. Huberman. 1994. *Qualitative Data Analysis* Thousand Oaks: SAGE.

MohammadSaif M., Bonnie J. Dorr, Graeme Hirst, and Peter D. Turney. 2013. "Computing Lexical Contrast." *Computational Linguistics* 39:555-90.

NanniAntonio and Mallory Fallin. 2021. "Earth, Wind, (Water), and Fire: Measuring Epistemic Boundaries in Climate Change Research." *Poetics* p. 101573.

NelsonLaura K.. 2020. "Computational Grounded Theory: A Methodological Framework." *Sociological Methods & Research* 49:3-42.

Nelson, Laura K. 2021. "Leveraging the Alignment Between Machine Learning and Intersectionality: Using Word Embeddings to Measure Intersectional Experiences of the Nineteenth Century U.S. South." *Poetics* p. 101539.

NelsonLaura K., Derek Burk, Marcel Knudsen, and Leslie McCall. 2021. "The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods." *Sociological Methods & Research* 50:202-37.

The New York Times. 1965. "Soviet Show Picketed in Ohio." *The New York Times*.

The New York Times. 1967. "Chicago Unit Sues to Fight Pollution of Lake Michigan." *The New York Times*.

The New York Times. 1972. "G.E. Resists War Protest; Honeywell Bars Arms Halt." *The New York Times*.

NothmanJoel, Hanmin Qin, and Roman Yurchak. 2018. "Stop Word Lists in Free Open-Source Software Packages." Pp. 7-12 in *Proceedings of Workshop for NLP-OSS*.

Ornaghi, Arianna, Elliott Ash, and Daniel L. Chen. 2019. "Stereotypes in High-Stakes Decisions: Evidence From US Circuit Courts." *Center for Law & Economics Working Paper Series* 2. doi:10.3929/ethz-b-000376877

Osgood, Charles Egerton, George J. Suci, and Percy H. Tannenbaum. 1957. *The Measurement of Meaning*. Champaign, IL: University of Illinois Press.

PangBo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. "Thumbs Up?: Sentiment Classification Using Machine Learning Techniques." Pp. 79-86 in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*.

Pardo-GuerraJuan Pablo and Prithviraj Pahwa. 2022. "The Extended Computational Case Method: A Framework for Research Design." *Sociological Methods & Research* 51:1826-67.

PaxtonPamela, Kristopher Velasco, and Robert W. Ressler. 2020. "Does Use of Emotion Increase Donations and Volunteers for Nonprofits?" *American Sociological Review* 85:1051-83.

PenningtonJeffrey, Richard Socher, and Christopher D. Manning. 2014. "Glove: Global Vectors for Word Representation." Pp. 1532-543 in *Proceedings of the 2014 Conference on EMNLP*.

PiantadosiSteven T.. 2014. "Zipf's Word Frequency Law in Natural Language." *Psychonomic Bulletin & Review* 21:1112-30.

RheaultLudovic and Christopher Cochrane. 2019. "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora." Pp. 1-22 *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association*.

RichieRussell, Wanling Zou, and Sudeep Bhatia. 2019. "Distributional Semantic Representations Predict High-Level Human Judgment in Seven Diverse Behavioral Domains." Pp. 2654-660 in *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.

Rinker, Tyler. 2022. "Lexicon: R Package." https://CRAN.R-project.org/package=lexicon.

RobertoFranzosi. 2021. "What's in a Text? Bridging the Gap Between Quality and Quantity in the Digital Era." *Quality & Quantity* 55:1513-40.

RodmanEmma. 2020. "A Timely Intervention: Tracking the Changing Meanings of Political Concepts With Word Vectors." *Political Analysis* 28:87-111.

Rodriguez, Pedro and Arthur Spirling. 2021. "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research." *Journal of Politics.* 84(1):101-115. 10.1086/715162

Rodriguez, Pedro, Arthur Spirling, and Brandon Stewart. 2023. "Embedding Regression: Models for Context-Specific Description and Inference." *American Political Science Review* 117(4):1255-1274. 10.1017/S0003055422001228

RollerStephen and Katrin Erk. 2016. "Relations Such as Hypernymy: Identifying and Exploiting Hearst Patterns in Distributional Vectors for Lexical Entailment." *arXiv* 1605.05433.

RossielloGaetano, Pierpaolo Basile, and Giovanni Semeraro. 2017. "Centroid-Based Text Summarization Through Compositionality of Word Embeddings." Pp. 12-21 in *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*. Association for Computational Linguistics.

RubnerYossi, Carlo Tomasi, and Leonidas J. Guibas. 1998. "A Metric for Distributions With Applications to Image Databases." Pp. 59-66 in *Sixth International Conference on Computer Vision*. IEEE.

SaltonGerard and Christopher Buckley. 1988. "Term-Weighting Approaches in Automatic Text Retrieval." *Information Processing & Management* 24:513-23.

SchlenderThalea and Gerasimos Spanakis. 2020. "'Thy Algorithm Shalt Not Bear False Witness': An Evaluation of Multiclass Debiasing Methods on Word Embeddings." *arXiv* 2010.16228.

SchmidtBenjamin. 2015. "Rejecting the Gender Binary: A Vector-Space Operation." http://bookworm.benschmidt.org/posts/2015-10-30-rejecting-the-gender-binary.html. Accessed: 2021-7-13.

SelivanovDmitriy, Manuel Bickel, and Qing Wang. 2020. "text2vec: Modern Text Mining Framework for R.".

Selivanov, Dmitriy, Manuel Bickel, and Qing Wang. 2020. "text2vec: Modern Text Mining Framework for R." https://CRAN.R-project.org/package=text2vec.

SiaSuzanna, Ayush Dalmia, and Sabrina J. Mielke. 2020. "Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics Too!" *arXiv* 2004.14914.

SidorovGrigori, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto. 2014. "Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model." *Computación y Sistemas* 18:491-504.

Snefjella, Bryor and Victor Kuperman. 2015. "Concreteness and Psychological Distance in Natural Language Use." *Psychological Science* 26(9):1449-1460. 10.1177/0956797615591771

StoltzDustin S. and Marshall A. Taylor. 2019. "Concept Mover's Distance: Measuring Concept Engagement Via Word Embeddings in Texts." *Journal of Computational Social Science* 2:293-313.

Stoltz, Dustin S. and Marshall A. Taylor. 2021. "Cultural Cartography With Word Embeddings." *Poetics* 80:101567. 10.1016/j.poetic.2021.101567

Stoltz, Dustin S, and Marshall A. Taylor. 2022. "text2map: R Tools for Text Matrices." *Journal of Open Source Software* 7(72):3741. 10.21105/joss

Stoltz, Dustin S., Marshall A. Taylor, and Jennifer S. K. Dudley. 2023a. "The Dynamics of Collective Action Corpus [Data set]." https://doi.org/10.5281/zenodo.8415049.

Stoltz, Dustin S., Marshall A. Taylor, and Jennifer S. K. Dudley. 2023b "Replication Repository for 'A Tool Kit for Relation Induction in Text Analysis'." https://doi.org/10.5281/zenodo.8415049.

Strauss, Anselm L.. 1987. *Qualitative Analysis for Social Scientists* Cambridge: Cambridge University Press.

TausczikYla R. and James W. Pennebaker. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29:24-54.

TaylorMarshall A. and Dustin S. Stoltz. 2020a. "Concept Class Analysis: A Method for Identifying Cultural Schemas in Texts." *Sociological Science* 7:544-69.

Taylor, Marshall A. and Dustin S. Stoltz. 2020b. "Integrating Semantic Directions With Concept Mover's Distance to Measure Binary Concept Engagement." *Journal of Computational Social Science* 4:231-242. 10.1007/s42001-020-00075-8

UtsumiAkira. 2020. "Exploring What Is Encoded in Distributional Word Vectors: A Neurobiologically Motivated Analysis." *Cognitive Science* 44:e12844.

ValentiniFrancisco, Germán Rosati, Diego Fernandez Slezak, and Edgar Altszyler. 2023. "The Undesirable Dependence on Frequency of Gender Bias Metrics Based on Word Embeddings.".

van DongenStijn and Anton J. Enright. 2012. "Metric Distances Derived From Cosine Similarity and Pearson and Spearman Correlations." *arXiv 1208.3145*.

van LoonAustin, Salvatore Giorgi, Robb Willer, and Johannes Eichstaedt. 2022. "Negative Associations in Word Embeddings Predict Anti-Black Bias Across Regions—But Only Via Name Frequency." *Proceedings of the International AAAI Conference on Web and Social Media* 16:1419-24.

Voyer, Andrea, Zachary D. Kline, and Madison Danton. 2022a. "Symbols of Class: A Computational Analysis of Class Distinction-Making Through Etiquette, 1922-2017." *Poetics* p. 101734.

Voyer, Andrea, Zachary D. Kline, Madison Danton, and Tatiana Volkova. 2022b. "From Strange to Normal: Computational Approaches to Examining Immigrant Incorporation Through Shifts in the Mainstream." *Sociological Methods & Research* 51(4):1540-1579.

VylomovaEkaterina, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2016. "Take and Took, Gaggle and Goose, Book and Read: Evaluating the Utility of Vector Differences for Lexical Relation Learning." Pp. 1671-682 in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

WangDan J. and Sarah A. Soule. 2012. "Social Movement Organizational Collaboration: Networks of Learning and the Diffusion of Protest Tactics, 1960-1995." *American Journal of Sociology* 117:1674-722.

WietingJohn, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. "Charagram: Embedding Words and Sentences via Character n-Grams." *arXiv 1607.02789*.

WilsonDavid S.. 1988. "Deaf Actress's Use of Speech Proves Divisive Among Peers." *The New York Times*.

Wood, Michael Lee. 2023. "Measuring Cultural Diversity in Text With Word Counts." *Social Psychology Quarterly*. Online First. 10.1177/01902725231194356

YıldırımSavaş and Tuğba Yıldız. 2018. "Learning Turkish Hypernymy Using Word Embeddings." *International Journal of Computational Intelligence Systems* 11:371-83.

YuShuiyuan, Chunshan Xu, and Haitao Liu. 2018. "Zipf's Law in 50 Languages." *arXiv* 1807.01855.

ZhangXuchao, Bo Zong, Wei Cheng, Jingchao Ni, Yanchi Liu, and Haifeng Chen. 2021. "Unsupervised Concept Representation Learning for Length-Varying Text Similarity." Pp. 5611-620 in *Proceedings of the 2021 Conference of the NAACL*.

ZhelezniakVitalii, Aleksandar Savkov, April Shen, and Nils Y. Hammerla. 2019. "Correlation Coefficients and Semantic Textual Similarity." *arXiv 1905.07790*.

Author Biographies

Dustin S. Stoltz is Assistant Professor of Sociology and Cognitive Science at Lehigh University. With Marshall A. Taylor, he is the author of the book *Mapping Texts: Computational Text Analysis for the Social Sciences* (Oxford University Press, 2024). His work has been published in peer-reviewed outlets such as *Sociological Theory, Poetics, Sociological Forum, Journal for the Theory of Social Behaviour, Journal of Computational Social Science,* and others.

Marshall A. Taylor is Assistant Professor of Sociology at New Mexico State University. His research revolves around questions of cognition and measurement in the sociology of culture. He is author, with Dustin S. Stoltz, of *Mapping Texts: Computational Text Analysis for the Social Sciences* (Oxford University Press, 2024). His work has been published in peer-reviewed outlets such as *Sociological Theory, Poetics, Political Behavior, Sociological Methods & Research, Journal of Computational Social Science,* and others.

Jennifer S. K. Dudley is a Postdoctoral Research Scholar in the Management Division at Columbia Business School. Her research focuses on inequality in social evaluations and group membership. Her work has been published in the *Journal of Evidence-Based Social* Work.