Voluntas (2023) 34:39–51
https://doi.org/10.1007/s11266-021-00399-7

RESEARCH PAPERS

Identifying Nonprofits by Scaling Mission and Activity with Word Embedding

Haohan Chen[1] · **Ruodan Zhang**[2]

Accepted: 18 August 2021 / Published online: 10 September 2021

Abstract This study develops a new text-as-data method for organization identification, based on word embedding. We introduce and apply the method to identify identity-based nonprofit organizations, using the U.S. nonprofits' mission and activity information reported in the IRS Form 990s in 2010–2016. Our results show that such method is simple but versatile. It complements the existing dictionary-based approaches and supervised machine learning methods for classification purposes and generates a reliable continuous measure of document-to-keyword relevance. Our approach provides a nonbinary alternative for nonprofit big data analyses. Using word embedding, researchers are able to identify organizations of interest, track possible changes over time and capture nonprofits' multi-dimensionality.

Keywords Nonprofit organizations · Text-as-data · Word embedding · Document retrieval · Identification

Introduction

The application of automated text-as-data methods to statements of mission and activity from official reports, websites, and tax forms has garnered growing interest in recent scholarship (Fyall et al., 2018; Litofcenko et al., 2020; Ma, 2021). Particularly, to study the U.S. nonprofit sector, the existing classification system developed by the National Center for Charitable Statistics and later adopted by the U.S. Internal Revenue Service (IRS), called the National Taxonomy of Exempt Entities (NTEE), falls short of meeting research demands to systematically capture nonprofits by type or service field. For example, the use of the NTEE codes in large-N studies often leads to misclassification, static measurement, or failure to facilitate more fine-grained analyses on a particular area (Fyall et al., 2018; Grønbjerg & Paarlberg, 2002). The release of the electronically filed tax form (Form 990s) data by the IRS offers new opportunities for nonprofit scholars to "remap" the sector (Ma, 2021) and to generate new research questions through examining the use of language (Messamore & Paxton, 2020).

Our research seeks to complement the existing application of machine learning classifiers by providing a nonbinary alternative to reimagine nonprofit taxonomy. We introduce and demonstrate a text-as-data method for measurement based on word embedding. Specifically, word embedding is a type of low-dimensional numeric representation of text data revealing the semantic relationship among words and documents. It serves our need to identify and retrieve documents, instead of classification. It generates a continuous measure based on organizational mission and program service activity statements. As such, it has unique advantages in capturing organizational mission changes over time, reflecting their multi-dimensionality, and identifying those in niche areas.

In the remainder of this paper, we first discuss the motivation to develop a continuous measure for organization identification. Next, we explain the intuition behind the word embedding algorithm and apply it to the mission and program service activity statements text data from the IRS Form 990s between 2010 and 2016. In the Results

✉ Ruodan Zhang
ruodan.zhang@uconn.edu

[1] Department of Politics and Public Administration, The University of Hong Kong, Hong Kong, China

[2] Department of Public Policy, University of Connecticut Hartford, 10 Prospect Street, Hartford, CT 06103, USA

section, we highlight the features that complement the existing automated methods for classification: time variance, measuring multi-dimensionality, and versatility. We conclude by discussing the limitations of this unsupervised method and call for methodological innovation in nonprofit research that enables researchers to examine new questions.

A Method for Identification

Organization classification is essential when nonprofit researchers seek to identify a sample of nonprofits by activity areas (e.g., Fyall et al., 2018) or to control for domain characteristics (e.g., Guo, 2007). In the U.S. context, the NTEE codes are the most widely used coding system. The classification system divides the sector into 26 major groups using alphabetic codes to cover all possible nonprofit activity areas. Within each group, organizations are further classified by numeric codes denoting the field of activity. Although adopted extensively, the coding scheme has been heavily criticized for various drawbacks. Ma (2021) summarizes five major problems of the NTEE codes, including misclassification due to organizations' multi-dimensionality; a static representation of potentially changing organizational missions and activities; high labor costs to update and manage the classification system; and finally, not covering organizations that do not file tax returns. Additionally, created to highlight nonprofit purposes and service areas (Barman, 2013), the NTEE system limits researchers' ability to explore the sector on other dimensions, such as mission motivation, core values, and distributive outcomes (Lecy et al., 2019).

Recent studies that advance nonprofit classification has used nonprofit mission and activity statements as the preferred classification resources (Brown, 2017; Fyall et al., 2018; Lecy et al., 2019; Litofcenko et al., 2020). A nonprofit's mission conveys its core intent and the target clients. It can be used by external stakeholders or regulatory agencies to hold the organization accountable (Fyall et al., 2018; Gugerty & Prakash, 2010; Lecy et al., 2019). Activities, or "program services" in the Form 990s, closely describe an organization's actual work to fulfill the stated intent (or the "exempt purposes"). Among the large-N studies, common approaches are dictionary methods (Fyall et al., 2018; Messamore & Paxton 2020) and supervised learning methods (e.g., Litofcenko et al., 2020; Ma, 2021). The dictionary approach utilizes researcher-developed dictionaries to search and code a sample, whereas the supervised learning approach relies on a correctly labeled "training" dataset and algorithms. In an evaluation of different automated text-as-data approaches, Plummer et al. (2019) argue that supervised learning may be more appropriate for classification than the dictionary approach, given its flexibility. Ma (2021) has demonstrated the reliability and accuracy of supervised learning; however, Litofcenko et al. (2020) show that supervised learning does require high-quality input text to begin with. Both studies used the existing classification systems to assess the method accuracy: i.e., the International Classification of Nonprofit Organizations (Salamon & Anheier, 1996) and the NTEE codes.

The two approaches for "nonprofit classification" essentially prioritize two sets of problems: one of identification and the other of classification. Specifically for nonprofit identification, as Fyall et al. (2018) and Messamore and Paxton (2020) have shown, it is not necessary to ground the process on any pre-determined structures (e.g., the NTEE system) or subscribe to their existing limitations. For instance, identity-based organizations (IBOs), which are nonprofits formed to benefit people with a shared identity, such as age, gender, religion, race/ethnicity, nationality, veteran status, and sexual orientation (Carvalho, 2016; Minkoff, 2002; Reid, 1999), may scatter over multiple NTEE major codes, but constitute an important group of organizations to be identified and studied systematically. Additionally, for niche areas, identification is more important than classification.

Therefore, the primary aim of this research is to suggest an identification method that complements the dictionary approach, by improving the flexibility and reducing the labor cost for dictionary generation. The method is based on inductive logic and minimum assumptions of expert knowledge and benefits from the flexibility of computer algorithms for text analysis.

Our second aim is to develop a continuous measure that is more informative than a binary assignment. Although the binary coding can indicate drastic mission drift across categories, a continuous measure captures to what extent a mission has departed from its previous versions, even when its primary area remains the same. Likewise, it can also help quantify the relative closeness between the text and a mission/concept.

Data and Methods

We obtain text information on nonprofit mission and program service activities from the electronically filed IRS Form 990s by 501(c)(3) nonprofit organizations between 2010 and 2016 (Applied Nonprofit Research, 2019a). On Form 990, a summary of mission and program activities is reported in Part I Line 1. We also collect mission statements from Part III Line 1 and program service descriptions from Part III Lines 4a-4c. For comparison, we use the NTEE codes compiled in a packaged nonprofit governance

data set (Applied Nonprofit Research, 2019b). Each organization is identified by the IRS Employer Identification Number (EIN). For repeated entries within the same tax period, we keep the latest submission for each organization. We concatenate the text information and exclude entries less than three words or those that contain expressions such as "none," "n/a," or "see attached." For data preprocessing, we convert letters to lowercase.

In short, we apply word embedding (or "distributed semantics"), a technique of Natural Language Processing that represents the semantics of words and documents with low-dimensional vectors (e.g., Mikolov et al., 2013).[1] Our full workflow starts with inputting the raw text data from Form 990s to train word embedding models with *word2vec*; next, we use the trained word vectors to embed query terms and documents, respectively; finally, we score the relevance of documents to the query terms of interest using a conventional cosine similarity measure. This process can later be combined with researchers' manual selection and interpretation for more precise nonprofit identification. This section explains the assumptions for word embedding and *word2vec*, the process to embed query terms and documents, and the matching process.

Word embedding. Word embedding is based on a simple linguistic idea called the *distributional hypothesis*: Words that occur in similar contexts have similar meanings (e.g., Joos, 1950; reviewed by Jurafsky & Martin, 2019, Chapter 6). The distributional hypothesis makes three assumptions. First, contexts are local, thus making it possible to infer the meaning of words using all other words around it as the context. Second, the relative relationships between words can be inferred without understanding the word's absolute meanings. Third, learning with the distributional hypothesis needs abundant text data.

Word embedding methods follow the distributional hypothesis and represent words with vectors of numeric values that best "describe" their linguistic context, which in turn represent the relative meanings. First, words with similar syntactic or topic meanings have word vectors numerically close to one another. For example, the word vector for *Christian* ($\mathbf{v}_{\text{Christian}}$) is numerically close to that of *Catholic* ($\mathbf{v}_{\text{Catholic}}$) or *Islamic* ($\mathbf{v}_{\text{Islamic}}$). ($\mathbf{v}_w$ denotes the word vector of a word w.) Second, the analogy of words can be captured by simple linear operations between word vectors (Mikolov et al., 2013). For example, $\mathbf{v}_{\text{Christian}} - \mathbf{v}_{\text{church}} \approx \mathbf{v}_{\text{Islamic}} - \mathbf{v}_{\text{mosque}}$ represents that "church" is to "Christian" what "mosque" is to "Islamic." Third, adding vectors of words in a phrase results in a valid representation of the phrase (

$$\mathbf{v}_{\text{Christian}} + \mathbf{v}_{\text{church}} \approx \mathbf{v}_{\text{Pentecostal}} + \mathbf{v}_{\text{church}}$$

). A weighted sum of word vectors in a document makes a valid representation of the document (Iyyer et al., 2015; Mitchell & Lapata, 2010; Arora et al., 2017).

The machine learning process based on word embedding assigns words with numeric vectors so that mathematical operations among these vectors can best explain "which words likely appear with which words in what context," given the text data. To obtain the "trained" word vectors, we use *word2vec*, a popular foundational prediction-based word embedding algorithm, to encode words in the text data from both mission and program activity statements (Mikolov et al., 2013; for further details, see Appendix A). We used all the text data for the benefit of a larger corpus.

Embedding query terms. Next, we decode query terms of interest by first looking up their word vectors and then finding their similar terms for validation and corpus augmentation. For method demonstration, we focus on retrieving IBOs, specifically, those of religion, race/ethnicity, and sexual orientation, none of which are readily identifiable using the NTEE codes. Each domain differs in the availability of existing NTEE codes, and in each case, using the NTEE codes alone would lead to errors of over- or under-counting. We attempt multiple query terms to retrieve their word vectors. We use "religious," "faith," "God," and "Buddhist" for religious identities, "Black," "African," "Asian," "Hispanic," and "Latino" for racial/ethnic identities, and "LGBT" and "gay" for sexual-orientation-related identities.

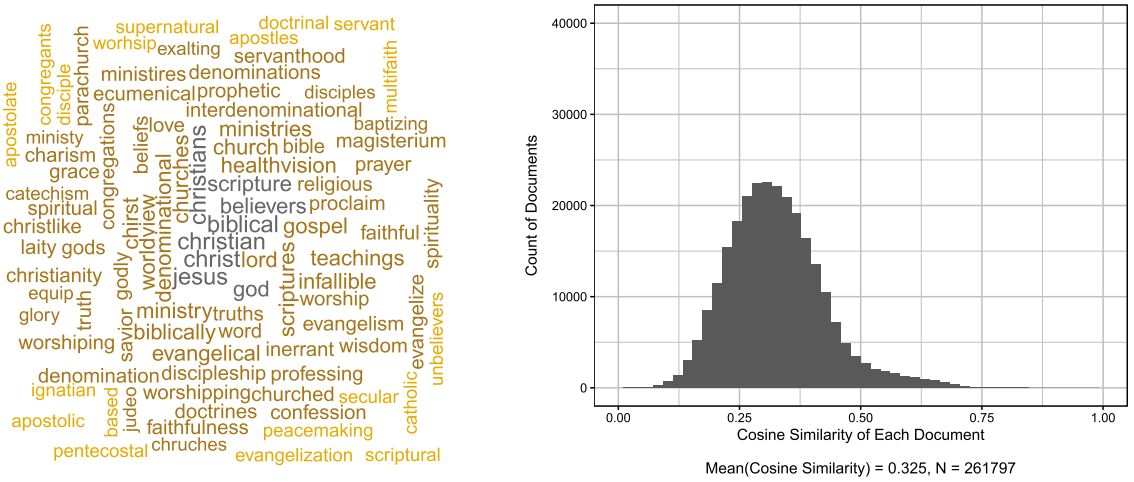
To make sense of the vectors, we rely on the property of distributed semantics: Words with similar functions and meanings are close to one another. We follow the existing literature to use cosine similarities as "measures of closeness":

$$\cos(\mathbf{v}_{w_i}, \mathbf{v}_{w_j}) = \frac{\mathbf{v}_{w_i}^T \mathbf{v}_{w_j}}{||\mathbf{v}_{w_i}|| \times ||\mathbf{v}_{w_j}||} \quad (1)$$

The cosine similarity method is typically used to evaluate word embedding algorithms as the measure is expected to capture generic language features of words.

With widely accepted "ground-truth" data, scholars may perform two types of standardized tasks for verification: word similarity/relatedness and word analogy (Levy et al., 2018, p. 217). In our application, assuming no "ground-truth" data, we follow our substantive knowledge to evaluate the quality of words whose vectors are close to the query terms.

[1] "Low-dimensional" numeric representation with word embedding turns every word into a 100–300-dimensional numeric "word vector." Word vectors capture the relationship among words, although their absolute values have no interpretable meanings. They are considered "low-dimensional" vectors relative to "high-dimensional" representation under previous methods, whose numeric word representations can take tens of thousands of dimensions.



(a) Word Cloud **(b)** Histogram of Cosine Similarity with Keyword "Faith"

Fig. 1 Descriptive statistics of example query terms "Faith." **a** Words highly relevant to the query term "Faith." Words with higher cosine similarity scores are located in the center. **b** Frequency distribution of the document cosine similarity scores (0–1) using the query term "Faith"

Embedding documents. Before matching our query terms to the documents, we embed documents in the same numeric space of distributed semantics. We use a "bag-of-word-vector" approach to embed documents based on learned distributed semantics of words (Arora et al., 2017). For a document, we take a weighted sum of distributed semantics of all words appearing in it. Formally, let D represent the collection of documents. Let d_i be a document in the text dataset, represented as a collection of words $d_i = \{w_1, w_2, ..., w_{M_i}\}$ where M_i is the number of unique words document d_i contains (the number of words may differ across documents). Let $\mathbf{v}(d_i)$ be the distributed semantics of document d_i. Let $f(d_i, w_j)$ be some weight assigned to word w_j in document d_i. Also, $\mathbf{v}_{w_j}$ is the distributed semantics of word w_j.

$$\mathbf{v}(d_i) = \sum_{w_j \in d_i} f(d_i, w_j)\mathbf{v}_{w_j} \tag{2}$$

Following common practice in bag-of-words methods, we create document vectors from the weighted average of word vectors, where the weights are the term-frequency-inverse document frequency (TF-IDF). Compared with a simple count, TF-IDF weighs down common words in the whole corpus and weighs up distinct words in individual documents. In this way, document vectors are subject to less influence of functional words like "purpose," "objectives," or "nonprofit," which frequently appear in our data.

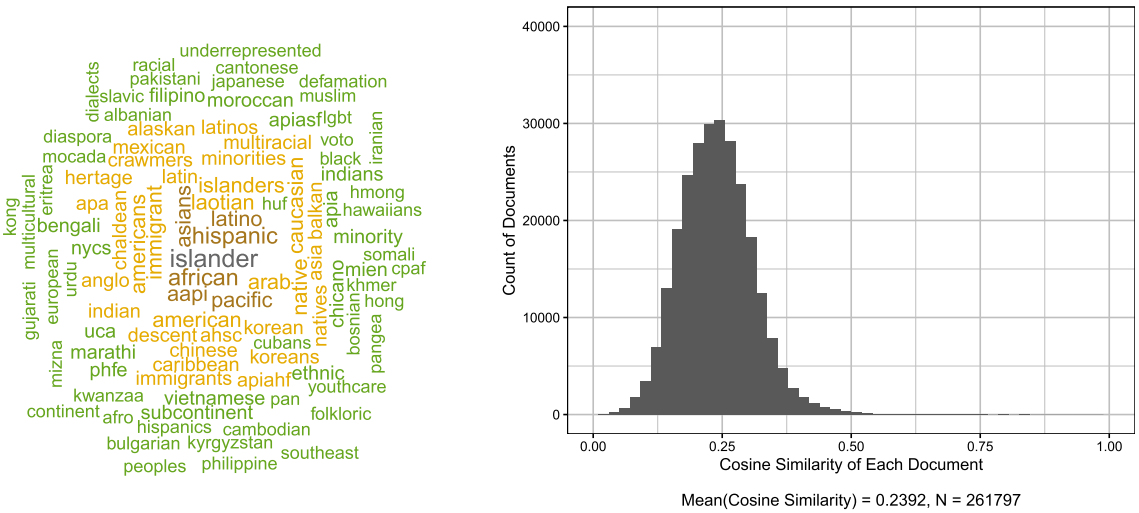
Taking the summation of word vectors and disregarding their order is justified by the compositionality property: The meanings of a phrase can be represented by the summation of vectors of words in the phrase. The authors of *word2vec* have demonstrated this important property (Mikolov et al., 2013, p. 7). Other natural language processing studies have empirically applied the compositionality property of word vectors to long documents (Mitchell & Lapata, 2010; Iyyer et al., 2015).

Admittedly, there are more complex word embedding algorithms for documents, such as the *doc2vec* model (Dai et al., 2015; Le & Mikolov, 2014) and the *Skip-thought* model (Kiros et al., 2015; Kim, 2014). However, our analyses show that the simple method suffices to retrieve documents close to the query terms of interest.

Matching query terms and documents. The final step matches the embedded query terms and documents. In the case of multiple query terms, we take simple or weighted averages of their word vectors. Formally, for a concept c_k:

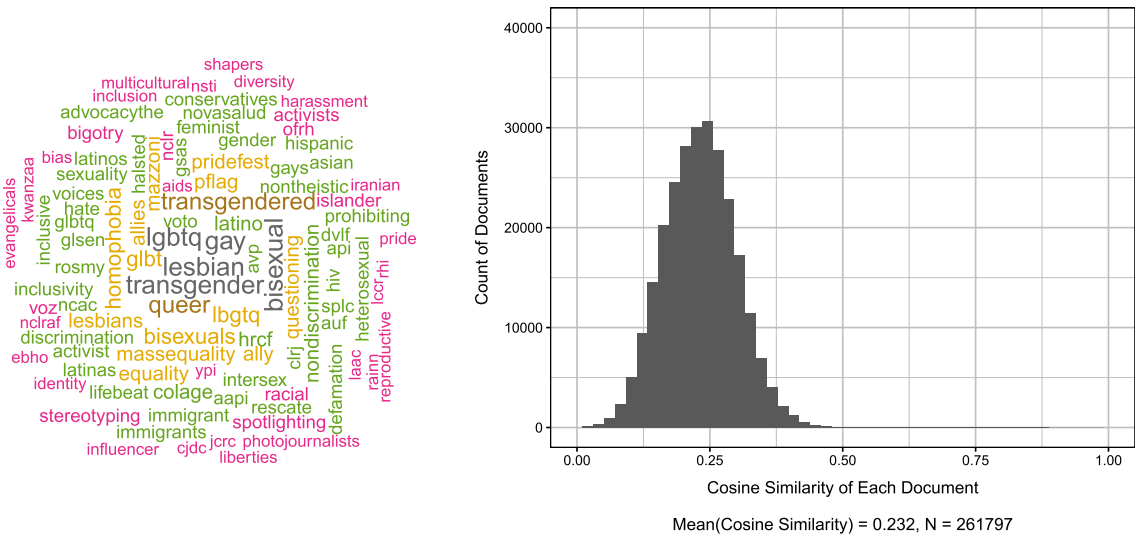
$$\mathbf{v}_{c_k} = \sum_{w_j \in c_k} \mathbf{v}_{w_j} \tag{3}$$

To show different levels of efficacy, we match the documents with single query terms. This allows us to compare the results across query terms.



(a) Word Cloud **(b)** Histogram of Cosine Similarity with Keyword "Asian"

Fig. 2 Descriptive statistics of example query terms "Asian." Notes: **a** Words highly relevant to the query term "Asian." Words with higher cosine similarity scores are located in the center. **b** Frequency distribution of the document cosine similarity scores (0–1) using the query term "Asian"



(a) Word Cloud **(b)** Histogram of Cosine Similarity with Keyword "LGBT"

Fig. 3 Descriptive statistics of example query terms "LGBT." Notes: **a** Words highly relevant to the query term "LGBT." Words with higher cosine similarity scores are located in the center. **b** Frequency distribution of the document cosine similarity scores (0–1) using the query term "LGBT"

The document-level cosine similarity score quantifies a document's relation with a concept by the cosine distance of their distributed semantics. Let $\text{score}_{d_i}^{c_k}$ be a score indicating the closeness between document d_i and concept c_k:

$$\text{score}_{d_i}^{c_k} = 1 - \cos(\mathbf{v}_{d_i}, \mathbf{v}_{c_k}) \quad (4)$$

Intuitively, this score is an indicator of the relevance between documents and query terms ranging from 0 to 1. A larger value of this indicator means the document is closer to the concept.

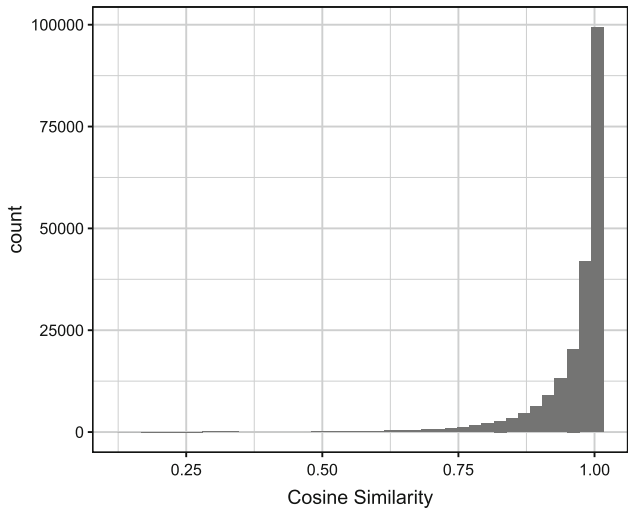


Fig. 4 Mission change (2010–2016) within organization

Results and Discussions

Summary Statistics

We summarize here results for three query terms of substantive interest. Figures 2, 3, and 4 show the descriptive information for the term "faith," "Asian," and "LGBT," respectively (see Appendix B and C for other terms used). Panel (a) of each figure highlights the words similar to the query term, as measured by cosine similarity scores. Panel (b) presents the distribution of cosine similarities between the query term and the documents. Figure 2b reveals a relatively higher overall salience of "faith" (mean = 0.3250) within the corpus. The right tail suggests that a significantly large number of nonprofits are closely related to the term. The salience of "Asian" is lower (mean = 0.2392), with a few outliers on the right ($x > 0.8$). It indicates that the outlying nonprofits are highly related to Asian ethnic identities whose statements are rather distinct. Finally, the salience of "LGBT" is close to that of "Asian" (mean = 0.2320).

Table 1 presents Pearson correlation coefficients between the query terms (for all values: $p < .0001$), which show three obvious clusters of "race/ethnicity," "religion," and "sexual orientation." We can also observe that within the race/ethnicity cluster, "black" reports the lowest correlation coefficients amongst all, due to its polysemy; within the religion cluster, "Buddhist" shares the least correlation with "faith" and "God," which are predominantly used in the Christian belief context. This is consistent with the word clouds. For example, we do not find "Buddhist" in Fig. 2a. Table 2 presents high similarity example documents of "faith," "Asian," and "LGBT." We found that although organizations may fall under various NTEE categories or provide different lengths of text, statements highly relevant to the term share similar scores.

The method is most efficient with either a broad or a highly specific term. We manually examined the mission and activity statements from the top 100 organizations in 2016 under each query term. All organizations are considered "highly relevant" to the query terms "faith" and "LGBT." Ninety-five out of 100 "Asian" organizations were strictly related to Asian identities or countries. The remaining five organizations have lower scores and serve either the African American or the Latinx community. With more ambiguous terms, such as "black," it is generally able to isolate racial/ethnic-related organizations because, within the nonprofit corpus, the word has more often been used in the racial/ethnic context. However, we also found 10 non-identity-based organizations (e.g., "Black Mountain College Museums," "Black Hawk College Foundation").

Table 1 Correlation coefficients between query terms

	African	Black	Hispanic	Latino	Asian	Religious	Faith	God	Buddhist	LGBT	Gay
African	1										
Black	0.7660	1									
Hispanic	0.7758	0.6387	1								
Latino	0.8237	0.6424	0.8978	1							
Asian	0.8690	0.7207	0.8390	0.8235	1						
Religious	0.3603	0.1933	0.395	0.3266	0.3591	1					
Faith	0.5351	0.2690	0.4873	0.5436	0.4495	0.7404	1				
God	0.4532	0.2179	0.3235	0.3741	0.3192	0.6545	0.904	1			
Buddhist	0.4503	0.4207	0.2791	0.2450	0.4294	0.6628	0.4743	0.4975	1		
LGBT	0.7363	0.6333	0.7341	0.8271	0.7767	0.3647	0.5353	0.3961	0.3318	1	
Gay	0.6623	0.6056	0.6371	0.7491	0.6743	0.3552	0.5097	0.3783	0.2982	0.9147	1

Table 2 Example match between query terms and form 990 statements

Query term	Example from 2016 Form 990 text
Faith	EIN: 510392520 Secular Institute of the Two Hearts (NTEE: X21), score = 0.75 Religious formation, seminars, conferences Evangelize and spread the teachings of JESUS CHRIST through Gospel readings/teachings, giving inspiring talks on the need to prepare oneself for ones call to discipleship, and the need to be devoted to educating oneself, sharing the faith and living as JESUS did. Develop youth spiritual, moral and leadership. Renew spiritual commitment and strengthen faith and promote spiritual unity among prayer parties
	EIN: 470446037 Nebraska Christian Schools (NTEE: B24), score = 0.74 The corporation operates a private Christian school to give children grades kindergarten through 12th grade a Christian education. To assist the family and church by providing a Christ-centered education, equipping students with a biblical worldview and encouraging a love relationship with the Lord Jesus Christ
Asian	EIN: 272577567 Asian Americans Advancing Justice Atlanta (NTEE: P84), score = 0.71. To protect and promote the civil, social, and economic rights of Asian Americans in the southeast through public policy, legal education, community organizing, and leadership development. Civil engagement; public policy work; legal services; legal education and outreach; legal advocacy for Asian American provide advocacy and education for Asian American legal understanding
	EIN: 133486145 National Asian-American Theatre Company (NTEE: not provided), score = 0.69. The organization presents the following repertory: European and American classics as written with all Asian American casts; adaptations of these classics by Asian American playwrights; and new plays—preferably world premieres written by non-Asian Americans, not for or about Asian Americans, but realized by an all Asian American cast. NAATCO asserts the presence and contributions of Asian American theatre artists in American culture by presenting theatre productions with all-Asian American casts in American and European classics, in adaptions of these classics by Asian American playwrights, and in works by non-Asian Americans
LGBT	EIN: 133850982 GLBT National Help Center (NTEE: S80), score = 0.77. Provide hotline services to the gay and lesbian bisexual and transgendered community. Services are completely free, and anonymous
	EIN: 561755564 Time Out Youth (NTEE: X20), score = 0.74. Youth Programs: Time Out Youth provides discussion groups-in 2015 193 youth accessed 48 groups. School Outreach: Time Out Youth provided 82 teacher trainings reaching 2,231 teachers. Emergency Housing and Emergency Financial Assistance-time out assists LGBTQ young adults who have been displaced from their homes for any reason by helping them become independent and self-sufficient individuals, to offer support, advocacy, and opportunities for personal development and social interaction to lesbian, gay, bisexual, transgender and questioning (LGBTQ) youth ages 11–20

Table 3 Example mission change and score: safe school certification

Year	Score	Mission/activity text from Part I Line I
2010	0.58	The organization works directly with lesbian, gay, bisexual, transgendered, and allied youth to cultivate advocates and leaders who fight homophobia and transphobia and strive for social justice
2013	0.57	IPN works directly with lesbian, gay, bisexual, transgendered, and allied youth to cultivate advocates and leaders who work to combat homophobia and trans-phobia and strive for social justice. Created and administers the Safe School Certification Program
2015	0.31	Administer the Safe School Certification Program , which leads Iowa schools to compliance with Iowa's anti-bullying legislation. Granted to One Iowa, an unrelated 501c3 organization
2016	0.26	Administer the Safe School Certification Program, which leads Iowa schools to compliance with Iowa anti-bullying legislation. This program was essentially inactive during this fiscal year, expecting to coordinate with a national organization in the near future to expand this program to other states. The organization incurred no program expense during the year. The organization has maintained a dormant state during the year, expecting to participate in a nationally funded program within the next year or two. This national program will be built upon and consistent with the programming developed by the organization

Method Versatility

Because the method primarily aims at identification, it complements other methods based on existing classification systems and is effective in identifying well-defined niche areas. Our decision to use the same corpus to train word embedding models makes the whole knowledge generation process a posteriori and completely based on how language has been used within this specific context.

Furthermore, since word embedding "learns" languages by the relational context, it can uncover underlying cultural associations (see also Kozlowski et al., 2019) and work

with multilingual corpora. For example, among the top "Latino" organizations, we found one with no mention of the query term in its mission/activity, except for one Spanish word: "Casa Ruby is the only Bilingual Multicultural LGBT Organization providing life-saving services and programs to the most vulnerable in the LGBT community. "

It is also possible to search by an n-gram (e.g., "climate change," or the whole mission statement of an organization) or a misspelled query term, as long as the misspellings also exist in the original corpus.

Mission Changes

Word embedding can be used to quantify changes in language use (Garg et al., 2018; Kozlowski et al., 2019). Garg et al. (2018) demonstrate how gender and racial/ethnic-related biases have evolved from 1910 to date, and the authors quantified such change over time using word embedding trained on Google Books, Corpus of Historical American English, and Google News. We compared the cosine similarities between the first and the last observation of an organization's mission and activities, and found the average similarity score was 0.966 (s.d. = 0.066), meaning that, overall, within the observations collected between 2010 and 2016, most nonprofits' mission/activities remained the same. Figure 4 shows that 34 percent of the nonprofits did not change their reported mission/activity statements at all (score = 1).

While comparing the "LGBT" lists with the NTEE-defined sample (P88 LGBT Centers and R26 lesbian and gay rights), we have identified an organization that experienced changes in both its name and mission in 2010–2016. Table 3 shows the mission text, in 2010, 2013, 2015, and 2016, respectively, changed from an identity-based framing to a program service framing. Correspondingly, the scores (on the query term "LGBT") dropped from 0.583 in 2010 to 0.261 in 2016.

Multi-dimensionality

Many nonprofit organizations work in multiple areas and represent a complex identity (Fyall et al., 2018; Ma, 2021). For example, *Hispanics in Philanthropy* is listed as a "general philanthropy, charity, and voluntarism promotion foundation" (T50). However, its mission indicates that it serves the Latino population and also forms "coalitions across the LGBT and Latino movements." We are able to rank all nonprofits in descending order of the similarity scores for a given query term, as a relative measure of relevance within the sample. Using its 2016 mission and activities, this organization ranks 293 under "LGBT," 478 under "gay," 71 under "Hispanic," and 3 under "Latino." For discriminant validity, we find it ranking 22,623 under "faith" and 87,520 under "religious."

The measure also reflects the level of commitment on each dimension. For example, both organizations below describe themselves as "religious." Yet the former has a significantly higher "religious" cosine similarity score, which reflects the textual differences:

> *Oklahoma City Youth for Christ* (score = 0.3726): Religious education of youth. Our organization offers an alternative group meeting to inner city students through Bible study and other Christian-related activities. We reach over 2,850 students per week. We also offer local churches an intern to help with their youth.
>
> *Juvenile Intervention and Faith-Based Follow-Up* (score = 0.6980): The purpose of the corporation is to give socially challenged and/or criminally oriented youth the skills, support, and direction necessary to break the cycle of destructive behavioral practices—enabling them to become thoughtful and productive citizens.

This could particularly be useful for nonprofit researchers interested in the study of intersectionality (Crenshaw, 1990) to explore relevant players in the sector.

Limitations

Although our proposed method proves reliable for identifying IBOs, we recognize a few limitations and would like to offer some recommendations. First, it is not intuitive how researchers may convert the continuous measure to a binary indicator (i.e., whether the organization is or is not related to a concept of interest), and there are no conventional "cutoff" thresholds. We suggest that researchers use human validation to determine the optimal threshold. Future studies may also compare results based on word embedding models with expert-developed dictionaries.

Second, training word embedding requires a large amount of text data. Unlike the dictionary approach, results from our method strictly rely on the size and variance of the training text and may require additional manual validation (Levy et al., 2018). To improve performance, we

recommend amending the current text dataset with web-scraped organizational information, which would greatly increase the richness of the data.

Finally, although we consider our proposed method reliable and cost-effective, future work can explore and integrate recent development to further improve document retrieval performance. For example, current state-of-the-art neural language models in natural language processing have adopted more complex representations of text data, such as representing words in context (Devlin et al., 2018; Liu et al., 2019).

Conclusion

In this paper, we propose a method for nonprofit identification based on the word embedding algorithm. We argue that this approach complements the existing methods to analyze nonprofit missions and activities. We apply the method to identify "identity-based organizations" and generate continuous cosine similarity scores. We show the method's potential to identify organizations of interest, including those in the niche areas, to quantify organizational changes over time and to capture organizations' multi-dimensionality.

The method does not require prior expert knowledge for nonprofit identification, and therefore, it allows the users to find a satisfactory list of organizations of interest with a relatively small and fuzzy set of keywords. Alternatively, researchers seeking higher precision may use the method only for generating a "seed dictionary" of the key concept/words and then amend or modify the seed dictionary using expert-generated information.

Word embedding is a powerful tool for nonprofit researchers and practitioners. Briefly, we highlight four new lines of research questions that can be facilitated by this method. First, nonprofit researchers may systematically identify and explore organizations serving small or emerging topical areas, such as climate change, opioid addiction support, or impact investing. It alleviates the "cold start" problem for nonprofit researchers examining new organization types without substantial expert knowledge.

Second, the method can be used to understand mission- or activity-related research questions. For example, how do nonprofits frame their missions and activities over time? What are the trends in program services provision for a mission area? We use both mission and activity statements to create organization embedding. However, as nonprofits tend to report more socially desirable goals and/or adopt similar issue frames (Plummer et al., 2019), we suggest that future research evaluates missions and activities separately to uncover possible discrepancies.

Third, researchers can use this method to obtain relative measures to understand how organizations serve multiple dimensions of identity or social needs. Some questions include: What are the closely connected mission/service areas? Where do the cross-identity collaborations happen? Do any intersectional disadvantages arise as an unintended consequence of nonprofits disproportionately serving particular communities?

Finally, word embedding has been applied to multilingual environments, for example, to study sentiments in YouTube comments (Nguyen & Le Nguyen, 2018). International or comparative studies researchers may explore how missions or service areas are framed in different cultural communities. For practitioners, the method can be applied to geo-coded data to identify similar nonprofits within a geographical area by N-gram searches.

Overall, word embedding shows strong promise in advancing the rigor and breadth of nonprofit research. More broadly, the method here is an example of how we can tap into text data to empirically explore nonprofit equity, justice, and value expressions across traditionally defined service fields. It is our hope that future efforts develop and apply word embedding models to solve new puzzles in the field. We encourage and join the call for innovative and boundary-spanning methods that involve different knowledge generation processes (*a priori* or *a posteriori*), generate nonbinary means to quantify outcomes or social constructs, and combine new types of data (e.g., text, geospatial, or audio/visual data).

Appendix A: Learning Distributed Semantics with *word2vec*

Word2vec has two variants: the Skip-gram model and the continuous bag-of-words (CBOW) model. The difference between the two lies in the specific task of prediction performed. The Skip-gram model predicts context words with the target word; the CBOW model predicts the target word with context words. Thus, the definition of likelihood functions slightly differs, leading to different optimization tasks.

To formalize, we start with the following setup: Consider a sequence of text of T words in total. Let the size of context windows be c (i.e., c's immediate neighbors before and after a word are considered its context). Let $\mathbf{v}(w)$ be the vector of distributed semantics of word w. Let the size of the vocabulary be V, (i.e., there are V unique words in the text). Let $p(w_i|w_j)$ be the probability of word w_i appearing, given word w_j. Let $\mathcal{L}$ be the likelihood.

The *word2vec* uses the a softmax function to link distributed representation of words (word vectors) with their predicted probabilities. Specifically, the probability of word w_i given word w_j in its context window is the exponential of the dot products of the word vectors $\mathbf{v}_{w_i}, \tilde{\mathbf{v}}_{w_j}$ over the sum of the exponentials of the dot products of $\tilde{\mathbf{v}}_{w_j}$ with word vectors of all words in the vocabulary:

$$\log p(w_i|w_j) = \log \frac{\exp(\mathbf{v}_{w_i}^T \tilde{\mathbf{v}}_{w_j})}{\sum_{k=1}^{V} \exp(\mathbf{v}_{w_k}^T \tilde{\mathbf{v}}_{w_j})} \tag{5}$$

$$= \mathbf{v}_{w_i}^T \tilde{\mathbf{v}}_{w_j} - \log \sum_{k=1}^{V} \exp(\mathbf{v}_{w_k}^T \tilde{\mathbf{v}}_{w_j}) \tag{6}$$

Thus, the log-likelihood of the Skip-gram model is computed as follows. At location t of the text sequence, the joint conditional probability of words in the context window (conditional on the target word at t) is calculated. The conditional probabilities are obtained by applications of softmax on the target word vector against each context word vector. Then the algorithm moves to location $t+1$ and repeat the process until the end of the sequence. The log-likelihood is the sum of all log probabilities. Formally:

$$\log \mathcal{L} = \sum_{t=1}^{T} \sum_{-c \le j \le c, j \ne 0} \log p(w_{t+j}|w_t) \tag{7}$$

$$= \sum_{t=1}^{T} \sum_{-c \le j \le c, j \ne 0} \log \frac{\exp(\mathbf{v}_{w_{t+j}}^T \tilde{\mathbf{v}}_{w_t})}{\sum_{k=1}^{V} \exp(\mathbf{v}_{w_k}^T \tilde{\mathbf{v}}_{w_t})} \tag{8}$$

$$= \sum_{t=1}^{T} \left[\sum_{-c \le j \le c, j \ne 0} \mathbf{v}_{w_{t+j}}^T \tilde{\mathbf{v}}_{w_t} - \log \sum_{k=1}^{V} \exp(\mathbf{v}_{w_k}^T \tilde{\mathbf{v}}_{w_t}) \right] \tag{9}$$

Similarly, the log-likelihood of the CBOW model is computed as follows: at location t of the text sequence, the probability of target word given context words is calculated. The conditional probability is obtained by a softmax of the target word vector and the average of context word vectors. Then the algorithm moves to location $t+1$ and repeats the process until the end of the sequence. The log-likelihood is the sum of all log probabilities. Formally:

$$\log \mathcal{L} = \sum_{t=1}^{T} \log p(w_t|w_{t-c}, w_{t-c+1}, \ldots w_{t+c-1}, w_{t+c}) \tag{10}$$

$$= \sum_{t=1}^{T} \log \frac{\exp(\mathbf{v}_{w_t}^T \bar{\mathbf{v}}_t)}{\sum_{k=1}^{V} \exp(\mathbf{v}_{w_k}^T \bar{\mathbf{v}}_t)} \tag{11}$$

$$= \sum_{t=1}^{T} \left[\mathbf{v}_{w_t}^T \bar{\mathbf{v}}_t - \log \sum_{k=1}^{V} \exp(\mathbf{v}_{w_k}^T \bar{\mathbf{v}}_t) \right] \tag{12}$$

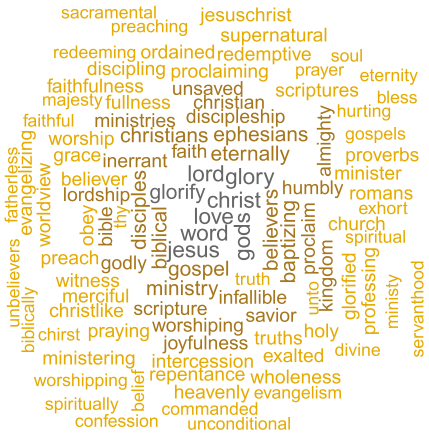
$$\text{where } \bar{v}_t = \frac{1}{2c} \sum_{-c \le j \le c, j \ne 0} \tilde{\mathbf{v}}_{w_{t+j}} \tag{13}$$

Both Skip-gram and CBOW models train vector representations of words to maximize the above defined likelihood. The processing is operationalized as neural networks trained by stochastic gradient descent. In general, they are both neural networks with one hidden layer and two weight matrices. The first weight matrices $\mathbf{W}_{V \times N}$ contain vector representations of all V words as targets in the vocabulary: $\mathbf{W}_{V \times N} = [\mathbf{v}_{w_1}, \mathbf{v}_{w_2}, \ldots, \mathbf{v}_{w_N}]^T$. The second weight matrices $\tilde{\mathbf{W}}_{N \times V}$ contain vectors of words as context: $\tilde{\mathbf{W}}_{N \times V} = [\tilde{\mathbf{v}}_{w_1}, \tilde{\mathbf{v}}_{w_2}, \ldots, \tilde{\mathbf{v}}_{w_N}]$. Input and output layers are one-hot-encoded words. The differences between Skip-gram and CBOW are evident in the model architectures. Skip-gram (Panel a) uses target words to predict context words, while CBOW (Panel b) uses context words to predict target words. Word vectors are updated with stochastic gradient descent. For the final output, researchers can use either of the two weight matrices $\mathbf{W}_{V \times N}, \tilde{\mathbf{W}}_{N \times V}^T$ or the two matrices' average as the representation of distributed semantics.

Training *word2vec* models can be computationally taxing. Two methods are used to reduce the computational demands of the model: hierarchical softmax and negative sampling. The algorithm in its naïve version described above can be computationally taxing primarily because the complexity of the softmax step (Eq. 5) grows linearly with the vocabulary size (i.e., $O(V)$ complexity): in the forward pass, it takes summations over the whole vocabulary of size V for the denominator; in the backpropagation, it updates all V word vectors in the vocabulary. Two methods have been developed to boost efficiency. First, hierarchical softmax uses a binary tree where words are represented by their leaf units. The probability of a word being the output is estimated by the probability of the path from root to leaf of the word. The method reduces computational complexity from $O(V)$ to $O(\log_2 V)$ given its tree structure. A second and more intuitive method, negative sampling, takes a random sample of words from the vocabulary to approximate the denominator in the forward pass and to update only the sample in the backpropagation. Thus, the computational complexity depends on the size of the negative sample and does not grow with the vocabulary size. The two methods have both demonstrated good performance in existing applications.

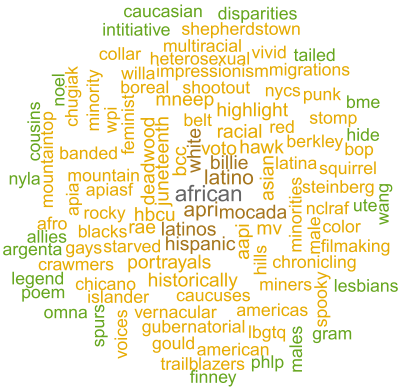
Appendix B: Word Cloud of Other Query Terms

(a) Gay



(b) God

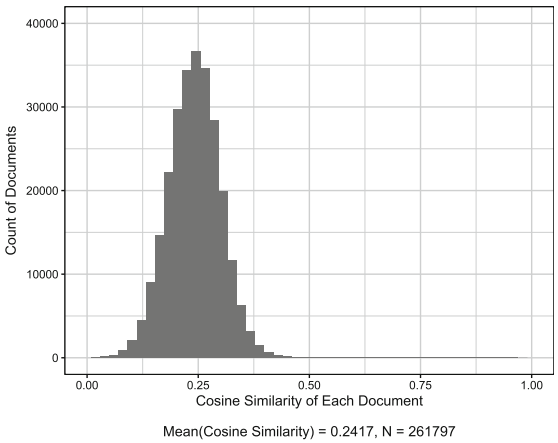
(c) Religious



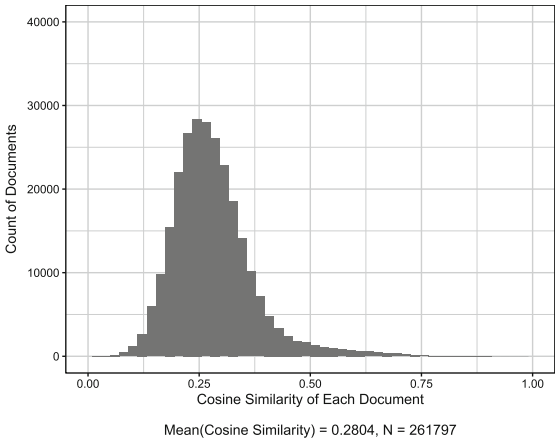
(d) Black

(e) Latino

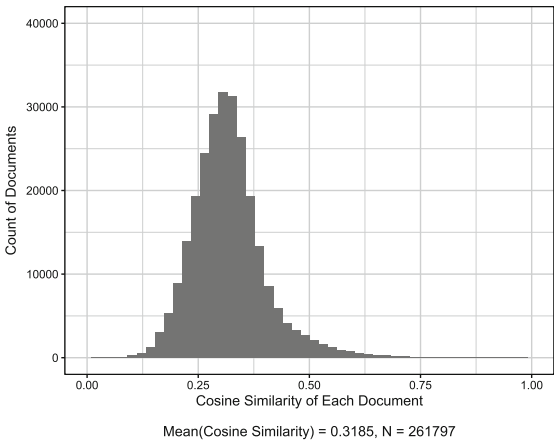
Appendix C: Additional Histograms of Cosine Similarity



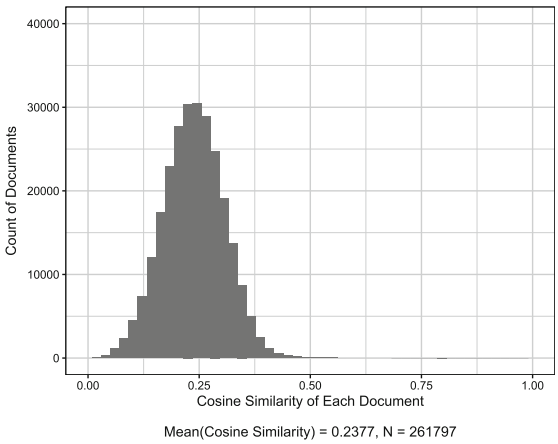
(a) Histogram of Cosine Similarity with Keyword "Gay"



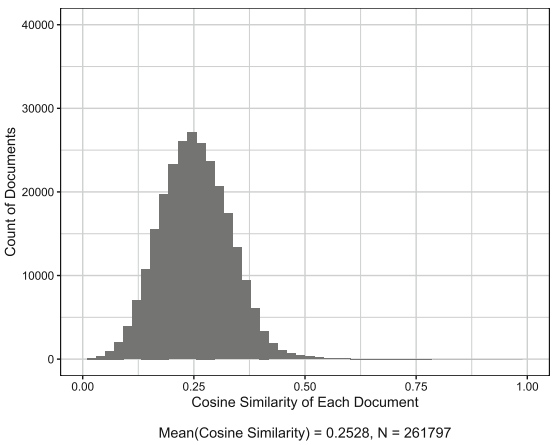
(b) Histogram of Cosine Similarity with Keyword "God"



(c) Histogram of Cosine Similarity with Keyword "Religious"



(d) Histogram of Cosine Similarity with Keyword "Black"



(e) Histogram of Cosine Similarity with Keyword "Latino"

Acknowledgments An earlier version of this paper was presented at the 2019 Association for Public Policy Analysis & Management Annual Conference. We thank the panel attendees, Yuan Cheng, the editors and anonymous reviewers for their constructive feedback; we thank Jonathan Richter for research assistance.

Author's Contribution Both authors contributed to the study conception and design. Material preparation and data collection were performed by RZ. Methodology and analysis were performed by HC. Both authors drafted, revised, read, and approved the manuscript.

Declaration

Conflict interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Applied Nonprofit Research (2019a). Form 990 variables (Versions 2009v1.4-2.12v3.0; 2013v3.0-2016v3.0).

Applied Nonprofit Research (2019b). Governance Dataset (Version 2019-01-15).

Arora, S., Liang, Y., & Ma, T. (2017). A simple but tough to beat baseline for sentence embeddings. In P*roceedings of International Conference on Learning Representations*.

Barman, E. (2013). Classificatory struggles in the nonprofit sector: the formation of the national taxonomy of exempt entities, 1969–1987. *Social Science History, 37*(1), 103–141.

Brown, W. (2017). Classification of program activities: How nonprofits create social value. *Administrative Sciences, 7*(2), 12.

Carvalho, J.-P. (2016). Identity-based organizations. *American Economic Review, 106*(5), 410–14.

Crenshaw, K. (1990). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review, 43,* 1241.

Dai, A. M., Olah, C., & Le, Q. V. (2015). Document embedding with paragraph vectors. arXiv preprint arXiv:1507.07998.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805

Fyall, R., Moore, M. K., & Gugerty, M. K. (2018). Beyond ntee codes: Opportunities to understand nonprofit activity through mission statement content coding. *Nonprofit and Voluntary Sector Quarterly, 47*(4), 677–701.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences, 115*(16), E3635–E3644.

Grønbjerg, K. A., & Paarlberg, L. (2002). Extent and nature of overlap between listings of irs tax-exempt registration and nonprofit incorporation: The case of Indiana. *Nonprofit and Voluntary Sector Quarterly, 31*(4), 565–594.

Gugerty, M. K., & Prakash, A. (2010). *Voluntary regulation of NGOs and nonprofits: An accountability club framework*. Cambridge University Press.

Guo, C. (2007). When government becomes the principal philanthropist: The effects of public funding on patterns of nonprofit governance. *Public Administration Review, 67*(3), 458–473.

Iyyer, M., Manjunatha, V., Boyd-Graber, J., & III H. D. (2015). Deep unordered composition rivals syntactic methods for text classification. In *Proceedings ofthe 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing* (pp. 1681–1691).

Joos, M. (1950). Description of language design. *The Journal of the Acoustical Society of America, 22*(6), 701–707.

Jurafsky, D. & Martin, J. H. (2019). *Speech and language processing: An introduction to natual language processing, computational linguistics, and speech recognition (3rd Edition)*.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1746–1751).

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., & Fidler, S. (2015). *Skip-thought vectors, 786,* 1–11.

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review, 84*(5), 905–949.

Le, Q. V. & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st international conference on machine learning* (Vol. 32).

Lecy, J. D., Ashley, S. R., & Santamarina, F. J. (2019). Do nonprofit missions vary by the political ideology of supporting communities? some preliminary results. *Public Performance and Management Review, 42*(1), 115–141.

Levy, O., Goldberg, Y., & Dagan, I. (2018). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics, 3,* 211–225.

Litofcenko, J., Karner, D., & Maier, F. (2020). Methods for classifying nonprofit organizations according to their field of activity: A report on semi-automated methods based on text. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations, 31*(1), 227–237.

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

Ma, J. (2021). Automated coding using machine learning and remapping the US nonprofit sector: A guide and benchmark. *Nonprofit and Voluntary Sector Quarterly*, *50*(3), 662–687.

Messamore, A., & Paxton, P. (2020). Surviving victimization: How service and advocacy organizations describe traumatic experiences, 1998–2016. *Social Currents*, 2329496520948198.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of NIPS, 2013,* 1–9.

Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, pp. 746–751.

Minkoff, D. C. (2002). The emergence of hybrid organizational forms: Combining identity-based service provision and political action. *Nonprofit and Voluntary Sector Quarterly, 31*(3), 377–401.

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science, 34*(8), 1388–1429.

Nguyen, H. T., & Le Nguyen, M. (2018). Multilingual opinion mining on youtube—A convolutional n-gram bilstm word embedding. *Information Processing and Management, 54*(3), 451–462.

Plummer, S., Hughes, M. M., & Smith, J. (2019). The challenges of organizational classification: A research note. Social. *Currents,* 2329496519878469.

Reid, E. J. (1999). *Nonprofit advocacy and political participation* (pp. 291–325). Nonprofits and government: Collaboration and conflict.

Salamon, L. M., & Anheier, H. K. (1996). *The international classification of nonprofit organizations*. Johns Hopkins University Institute for Policy Studies Baltimore Mar.