CLASSIFICATION OF NONPROFIT ORGANIZATIONS: A SUPERVISED MACHINE-LEARNING APPROACH

JI MA AND ISHA KANANI [*]

University of Texas at Austin

April 24, 2019

Abstract

This research note reports the use of supervised machine-learning algorithms in classifying the nonprofit organizations in the United States. Mission statements and project descriptions are collected from the 990 forms as text data, and classifications using National Taxonomy of Exempt Entities are collected from the National Center for Charitable Statistics at the Urban Institute. Three text classification algorithms are experimented: Naïve Bayes, Random Forest, and Neural Network. The Neural Network classification achieves the best results with an average accuracy of 9*.9% (standard deviation **), recall *** (standard deviation **), and precision *** (SD **). An open-source Python package *npocat* is developed and shared using the trained algorithms. Future projects are discussed.

[*] J.M.: maji@austin.utexas.edu, LBJ School of Public Affairs and RGK Center for Philanthropy and Community Service; I.K.: ishakanani@utexas.edu, School of Information.

Contents

List of Tables

List of Figures

1 Introduction

Although the voluntary and philanthropic organizations have long been existent for numerous centuries, the so-called "nonprofit sector" was only coined in the 1970s by scholars, policy makers, and nonprofit practitioners. A major reason for assembling the diverse organizations as a conceptual whole is to legitimize the existence of these organizations and the benefits these organizations receive (Hall, 2006, pp. 54–55; Barman, 2013). From Durkheim's 2012 perspective, the order and structure of a society can be reflected by a classification system. The National Taxonomy of Exempt Entities (NTEE) developed by the National Center for Charitable Statistics (NCCS), the most widely used classification system, is one of the efforts legitimizing the existence of nonprofit sector (Hodgkinson, 1990; Hodgkinson & Toppe, 1991). As Barman (2013, p. 105) cite Clarke and Casper (1996, p. 601): "The ways in which different entities (people, animals, plants, diseases, etc.) are organized into classificatory groups reveal something of the social, cultural, symbolic, and political contexts within which classifications occur."

The development of NTEE classifications can date back to the 1980s (Hodgkinson, 1990, pp. 8–9, 11). In 1982, NCCS assembled a team of experts working on creating a taxonomy for nonprofit organizations. The first draft of the taxonomy, entitled "National Taxonomy of Exempt Entities" (NTEE), came out in 1986 and published in 1987. In the early 1990s, NCCS had classified nearly one million nonprofits using NTEE. In 1995, the Internal Revenue Service (IRS) adopted the NTEE coding system, took over the tasks of assigning and maintaining the classifications, and started to release the Business Master File with NTEE codes (US Internal Revenue Service, 2013, 2014).

Two agencies took the task of assigning NTEE codes: NCCS and IRS. Before 1995, NCCS coded nonprofits according to the program descriptions in Part III and VIII of Form 990, supplemented with information from Form 1023 ("Application for Recognition of Exemption") and additional research (National Center for Charitable Statistics, 2006, p. 16). After 1995, IRS began to issue "new exempt organizations an NTEE code as part of the determination process," and "the determination specialist assigns an NTEE code to each organization exempt under I.R.C. §501(a) as part of the process of closing a case when the organization is recognized as tax-exempt" (US Internal Revenue Service, 2013, p. 1).

The NTEE classifications has been used for numerous practical and academic purposes. For example, it provides a framework on which the social and economic activities of nonprofits can be mapped and compared with other types of organizations in a society (e.g., Roeger, Blackwood, & Pettijohn, 2015). It can also serve as an analytical tool for measuring the organizational capacity in different service domains and inform the practitioners and policymakers in decision-making (Hodgkinson & Toppe, 1991). Scholars also use NTEE codes for sampling purposes (e.g., Carman & Fredericks, 2010; Okten & Weisbrod, 2000) or as independent variables (Sloan, 2009). The invention of NTEE also provides a fundamental necessity for comparative international research, facilitating the study of "global civil society" (Hodgkinson, 1990; Lester M. Salamon & Anheier, 1992; Lester M Salamon, Anheir, & coaut, 1996; Vakil, 1997).

The NTEE classification system, although one of the best we have, still has several major drawbacks. First, because it only assign one major category code to an organization, it cannot accurately describe a nonprofit organization's programs which are usually diverse and across several domains (Grønbjerg, 1994,

p. 303). Although another classification system assigning purpose codes to programs was developed (Lampkin, Romeo, & Finnin, 2001), it is not widely used (why?). Second, the assignment of NTEE codes is not complete because it is "based on an assessment of program descriptions contained in Parts 3 and 8 of the Form 990" and "program descriptions were only available for some organizations" (National Center for Charitable Statistics, 2006, p. 16). A recent study found the number of organizations in Washington State with a specific NTEE code could be significantly increased if the mission statements were used for coding (Fyall, Moore, & Gugerty, 2018). Third, NTEE codes are static but nonprofit organizations' activities may change over time. Recoding existent NTEE assignments is extremely onerous, and this may be one of the reason that IRS does not have a procedure by which the nonprofits can request the change of their NTEE codes (US Internal Revenue Service, 2013).

NTEE limitations: Lester M. Salamon and Anheier (1992).

Contribution of this study.

2 Method

2.1 Working with Texts and Research Workflow

Classifying texts is a typical task of automatic content analysis and usually employs three types of methods: dictionary, supervised, and unsupervised methods (Grimmer & Stewart, 2013, pp. 268–269). The dictionary methods use a predefined dictionary of words to classifying the texts. Although accurate, this approach is not capable to deal with the variations and contexts of language. The supervised method is an improved solution which uses computer algorithms to "learn" the linguistic patters in a dataset classified by human coders. Unlike the dictionary and supervised methods which require predefined categories of interest, unsupervised methods can discover linguistic patters in texts without inputting any knowledge of classification. However, unsupervised method's validity can be problematic because the returned classifications may not be theoretically meaningful. To take the advantage of existing human-coded NTEE classifications, this study employs a supervised approach as indicated by Figure 1.

Figure 1 shows this paper's complete workflow. We implement four stages of analysis: 1) *preprocessing stage* includes data acquisition and the preprocessing of datasets and texts; 2) *feature extraction* includes bag-of-words (used by Naive Bayes and Random Forest algorithms) and word embedding (used by neural network algorithms); 3) at *training and intermediate decision making* phase, we use stochastic and grid search to train, search, and optimize the machine learning algorithms; 4) we *train the model finalist* with the complete dataset and prepare the trained model for public use. The following part introduces the four phases in detail.

2.2 Data Preprocessing

Data acquisition and dataset preprocessing. We collected text records from form 990, 990-EZ, and 990-PF, and supplemented these records with program descriptions from Schedule O. Form 990 (Return of Organization Exempt From Income Tax) is submitted by most of the nonprofit organizations. For smaller

Figure 1: RESEARCH WORKFLOW

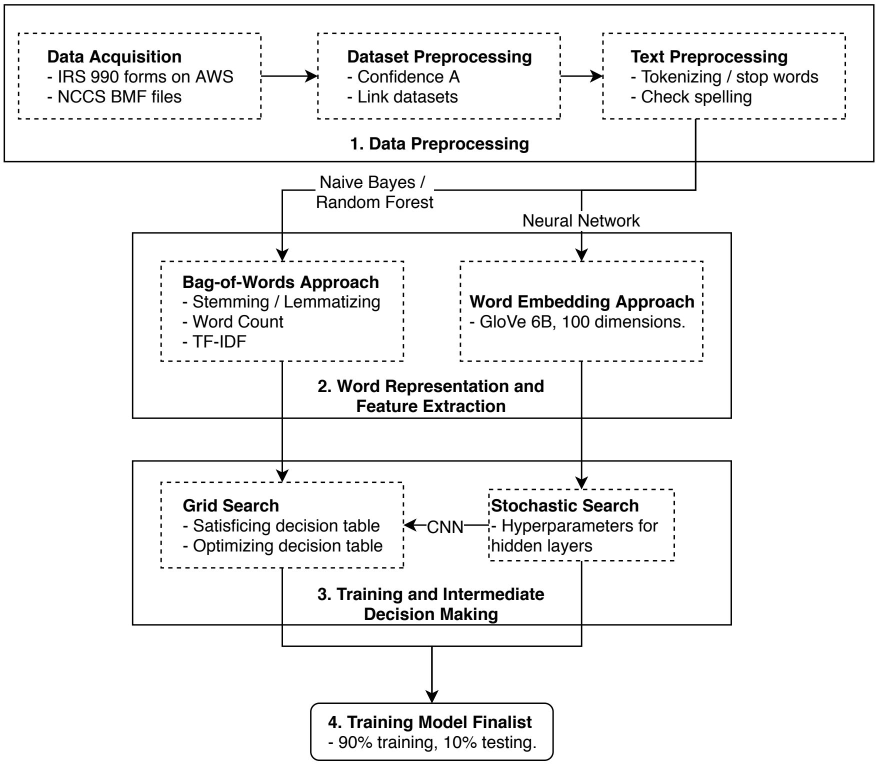
Data Acquisition
- IRS 990 forms on AWS
- NCCS BMF files
Dataset Preprocessing
- Confidence A
- Link datasets
Text Preprocessing
- Tokenizing / stop words
- Check spelling
1. Data Preprocessing
Naive Bayes / Random Forest
Neural Network
Bag-of-Words Approach
- Stemming / Lemmatizing
- Word Count
- TF-IDF
Word Embedding Approach
- GloVe 6B, 100 dimensions.
2. Word Representation and Feature Extraction
Grid Search
- Satisficing decision table
- Optimizing decision table
CNN
Stochastic Search
- Hyperparameters for hidden layers
3. Training and Intermediate Decision Making
4. Training Model Finalist
- 90% training, 10% testing.

Table 1: Locations of text fields in different forms.

	Mission Statement	Program Description
990	Part I, Line 1; Part III, Line 1	Part III, Line 4; Part VIII, Line 2a-e, Line 11a-c; Schedule O
990-EZ	Part III	Part III, Line 28-30; Schedule O
990-PF	–	Part IX-A; Part XVI-B

organizations with "gross receipts of less than $200,000 and total assets of less than $500,000 at the end of their tax year" (US Internal Revenue Service, 2018, p. 1), they can file Form 990-EZ (Short Form Return of Organization Exempt From Income Tax), a shorter version of Form 990. Private foundations use Form 990-PF (Return of Private Foundation). The texts describes organizational activities in two forms: overall mission statement and specific program description. Table 1 summarizes these text fields' specific locations in different forms.

Classification records (i.e., NTEE codes) are collected from the 2014-2016 Business Master Files on NCCS website.[1] The accuracy of classification is indicated by a letter of A, B, or C: "A confidence level of A, for example, indicates that there is at least a 90 percent probability that the major group classification is correct" (National Center for Charitable Statistics, 2006, p. 16). From 2014 to 2016, 56.12% records are classified at A level, 37.32% at B level, and 6.56% at C level. For training purposes, we only use records at confidence level A. About 1.76% organizations changed their NTEE codes between 2014 and 2016. We

[1] https://nccs-data.urban.org

Table 2: EXAMPLE OF COUNT VECTORS

statements X vocabulary	we	focus	on	education	health	care	about
we focus on education	1	1	1	0	0	0	0
health care care	0	0	0	0	1	2	0
we care about	1	0	0	0	0	1	1

dropped the records of these organizations since these records are less credible.

Text Preprocessing. Texts in sentences need to be "tokenized" into words before analysis, which is called "tokenization" in natural language processing. We also removed stop words (e.g., "the", "a" and punctuation marks) and checked spelling errors using algorithms based on "minimum edit distance" (ie., the minimum number of editing operations needed to change one word into another; Jurafsky and Martin (2017, p. 26)).

2.3 Word representation and feature extraction

The machine learning algorithms can only work on numeric vectors that are transformed from the tokenized sentences. A variety of transformation methods can "represent" words as vectors, and good methods should be able to easy the process of extracting "features" from texts. In general, there are two approaches to word representation: bag-of-words and word embedding.

2.3.1 Bag-of-words approach

Bag-of-words approach considers words in texts as mutually independent, as a result, disregards the order of words in text. For example, "we are health service organization" and "health organization service are we" are the same bag-of-words. This approach serves as the basis for developing many simple language models because it can efficiently represent the possibility of word's occurrence in texts (Bengfort, Bilbro, & Ojeda, 2018). We adopt two methods in this study to represent the texts: count vector and Term Frequency-Inverse Document Frequency.

Count vector counts the number of occurrences of all the words in a given text. Given a set of statements, the algorithm first builds an index of all unique words from the collection which is called vocabulary index. The algorithm then represent the texts using words' frequencies and vocabulary index. Table 2 presents a simple example of count vectors, in which "we focus on education" is represented as vector $[1,1,1,0,0,0,0]$

Term Frequency-Inverse Document Frequency (TF-IDF) normalizes raw word frequencies using the number of documents in which the word appears. As Eq. 1 presents, tf_{ij} is the frequency of word i in mission statement j, weighted by the inverse document frequency (i.e., idf_i; Eq. 2), where N^{total} is the number of total mission statements and N^i is the number of mission statements that word i appears. The underlying assumption of TF-IDF is that the words appear in all statements are not as important as those occur in a limited number of statements (Jurafsky & Martin, 2017, p. 278).

$$w_{ij} = tf_{ij} \cdot idf_i \tag{1}$$

$$idf_i = log(\frac{N^{total}}{N^i}) \tag{2}$$

We need to "normalize" the texts to reduce the vocabulary size before transforming using either count vector or TF-IDF, because the same word can have numerous spelling variations. For example, "environments," "environmental," and "environment" represent the same root word (i.e., *stem*) "environ." Otherwise, the ML models will suffer from "the curse of dimensionality": as the feature increases, the data becomes more discrete and less informative to decision making (Bellman, [1961] 2015, p. 94).

The process of finding stems is called "morphological parsing" which includes two primary methods: stemming and lemmatizing (Jurafsky & Martin, 2017, p. 25). Stemming slices longer strings to smaller ones according to a series of predefined rules. For example, "ational" is transformed to "ate" in all words ending with the former string. Therefore, stemming tend to have errors of both over- and under-parsing. Lemmatizing is a more advanced method which reduces a word to its stem by analyzing its meaning.

2.3.2 Word embedding approach

Disregarding the contexts in which the words appear is an evident drawback of bag-of-words approach. The word embedding approach represents words in a multi-dimensional space (i.e., each word has a vector value), in which words that often appear together in texts have closer distance with each other (Jurafsky & Martin, 2017, p. 290; Bengfort et al., 2018, p. 65). We can either train our own word vectors which require a large corpus and is time-consuming, or use pre-trained word vectors. We use the 100-dimension word vectors pre-trained from a corpus of 6 billion word tokens (Pennington, Socher, & Manning, 2014).

2.4 Training and intermediate decision making

2.4.1 Classifiers for Training

Naïve Bayes classifier is built on Bayeś theorem. It is one of the simplest classifiers to learn and implement among all machine learning algorithms and built on simple conditional probability principles. The classifier assumes all features extracted from the texts are conditionally independent, which is wrong in most cases. But the classifier is efficient and has proven to be useful for a variety of tasks even on a small dataset (Jurafsky & Martin, 2017, p. 76; Grimmer & Stewart, 2013, p. 277).

Random Forest classification is implemented by developing multiple prediction models. Each model in this algorithm is trained by different data, and then all of these models are asked to predict for the same record. A prediction class that is elected by most of these small algorithms is given as the prediction result by the random forest algorithm. It uses the word "forest" because each small algorithm trained is a decision tree (Quinlan, 1986, p. 83). A decision tree represents a set of questions that usually have Yes/No answers. The process starts from the top of the tree with one question, and based on the answer, we further run down on either one side of the tree, and answer another question and repeat till we reach the end of the tree. Each decision tree is trained on a different training set (Breiman, 1996, p. 124).

Take our study for example, if we provide 5,000 statements with their NTEE codes to a random forest with five trees, each tree will randomly select a thousand records to train. Each tree includes new words at different levels of the tree. For example, the tree starts with word “emergency,” it will have two branches for “yes” or “no,” leading to another word and so on, till the bottom of the tree where the NTEE code is. Since each tree is trained on a different set, when a record is given for prediction, each tree predicts the class independent of other tree. In total of five class predictions will be collected in this case, and the class which has the highest occurrence in the prediction results is given as the final predicted class by the random forest algorithm.

Since each decision tree in a Random Forest classifier is provided with a unique set of records for the training purpose, it strengthens the performance of the overall forest. The classifier however is difficult to visually interpret. It takes a little effort to visualize how decision trees work and understand the algorithm, unlike the Naïve Bayes approach.

Neural Network (NN) classification is built on the concepts of a neuron structure of the human mind. Each neuron in the network is connected to a few other neurons of the network by a numerical value called “weight.” The neurons process records each one in turn, and learn by looking at their classification (i.e., NTEE code in this case) with the known previous NTEE codes of records. With every new record the neurons learn, they update the connection value “weight” to update the model (Collobert & Weston, 2008, p. 163). After the network is done processing each record of the training set, it has final weights for each connection between two neurons. When a testing set is provided, the neurons use the final weights to predict the NTEE code. Depending on the architecture of the neurons, we can design a variety of NNs (e.g., the basic fully connected, Recurrent, and Long Short-Term Memory). We test the Convolutional NN (CNN) in this study (Zhang & Wallace, 2015).

NN algorithms beat many other machine learning algorithms in most cases. A significant amount of work available compares the performance of different approaches for the same dataset and neural networks algorithm beats many times. However, the model only works well when the data is in a significant amount. Along with the large size of data, this classification method also consumes a high computation power to train the network. One of the biggest disadvantages of Neural Networks is its “Black Box Nature” (Benitez, Castro, & Requena, 1997), which means that it is difficult to interpret the training process of this approach. There is no pre-defined algorithm that would provide the best results for all data format like text, audio, image etc. Also, the model yielding good accuracy for one text dataset might not give satisfactory results for another text dataset. One has to test different configurations to get the model that works best for the target dataset.

2.4.2 Measuring performance and intermediate decision making

Measuring algorithm performance. The performance of a classification algorithm can be measured by accuracy, precision, and recall. The *accuracy* measures the percentage of correctly classified organizations as showed in Eq. 3, where i is one of the three classification algorithms (i.e., NB, RF, and NN), $Org^{correct}$ is the number of organizations correctly classified by the algorithm i, and Org^{total} is the total organizations to be classified. For example, $Accuracy^{RF} = 0.6$ indicates that, when RF classifies an organization, the chance of getting right is 60%.

$$Accuracy^i = \frac{Org^{correct}}{Org^{total}} \tag{3}$$

The *precision* and *recall* measures the performance of a classifier on a specific category. In Eq. 4, k is one of the NTEE codes, $Org_k^{correct}$ is the number of organizations correctly classified as k by algorithm i, and Org_k^i is the number of organizations classified as k by algorithm i. $Org_k^{correct}$ will always be smaller than or equal to Org_k^i because ML algorithms can hardly predict everything right. For example, $Precision_B^{NN} = 0.75$ indicates that 75% of all the organizations classified as "education" by the NN algorithm are correct.

$$Precision_k^i = \frac{Org_k^{correct}}{Org_k^i} \tag{4}$$

Given a human coder labels an organization as category k, the *recall* measures the chance the classifier i also identifies the organization as k. In Eq. 5, Org_k^{hum} is the number of organizations that has been classified as k by human coders. For example, $Recall_B^{NN} = 0.80$ denotes that 80% of the organizations classified as "education" by human coders are correctly identified by the NN algorithm.

$$Recall_k^i = \frac{Org_k^{correct}}{Org_k^{hum}} \tag{5}$$

Intermediate decision making. Finding the best ML algorithm with appropriate parameters is the goal of this study. We can either try some of the configurations randomly (i.e., *stochastic search*), or iterate all possible configurations (i.e., *grid search*). For NB and RF algorithms, we used the latter approach. For NN algorithms, we first used stochastic search to narrow down the configurations of hidden layers, and then conducted a grid search for the input and output layers' parameters using CNN. The grid search for all possible parameter configurations (over 2 million combinations) is impossible by even using one of the most advanced super computing clusters in the world.

We conduced two rounds of grid search. The first found is for *satisficing decision making* in which we only considered the configurations that can perform at the top 5 percent. Then we ran the second found grid search for *optimizing decision making* in which we increased the values of some parameters to allow the algorithms to reach their performance ceilings. We then choose the best algorithm and parameters for final training.

3 Results

References

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.

Bellman, R. E. ([1961] 2015). *Adaptive Control Processes, A Guided Tour*. doi:10.1515/9781400874668

Bengfort, B., Bilbro, R., & Ojeda, T. (2018). *Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning* (1 edition). Beijing Boston Farnham Sebastopol Tokyo: O'Reilly Media.

Benitez, J. M., Castro, J. L., & Requena, I. (1997). Are artificial neural networks black boxes? *IEEE Transactions on Neural Networks*, *8*(5), 1156–1164. doi:10.1109/72.623216

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. doi:10.1007/BF00058655

Carman, J. G., & Fredericks, K. A. (2010). Evaluation Capacity and Nonprofit Organizations: Is the Glass Half-Empty or Half-Full? *American Journal of Evaluation*, *31*(1), 84–104. doi:10.1177/1098214009352361

Clarke, A. E., & Casper, M. J. (1996). From Simple Technology to Complex Arena: Classification of Pap Smears, 1917-90. *Medical Anthropology Quarterly*, *10*(4), 601–623.

Collobert, R., & Weston, J. (2008). A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 160–167). ICML '08. doi:10.1145/1390156.1390177

Durkheim, É. (2012). *The Elementary Forms of the Religious Life*. Courier Corporation.

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. doi:10.1177/0899764018768019

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. doi:10.1093/pan/mps028

Grønbjerg, K. A. (1994). Using NTEE to classify non-profit organisations: An assessment of human service and regional applications. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, *5*(3), 301–328. doi:10.1007/BF02354038

Hall, P. D. (2006). A Historical Overview of Philanthropy, Voluntary Associations, and Nonprofit Organizations in the United States, 1600–2000. In W. W. Powell & R. Steinberg (Eds.), *The nonprofit sector: A research handbook* (pp. 32–65). Yale University Press.

Hodgkinson, V. A. (1990). Mapping the non-profit sector in the United States: Implications for research. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, *1*(2), 6–32. doi:10.1007/BF01397436

Hodgkinson, V. A., & Toppe, C. (1991). A new research and planning tool for managers: The national taxonomy of exempt entities. *Nonprofit Management and Leadership*, *1*(4), 403–414. doi:10.1002/nml.4130010410

Jurafsky, D., & Martin, J. H. (2017). *Speech and Language Processing* (3rd draft).

Lampkin, L., Romeo, S., & Finnin, E. (2001). Introducing the Nonprofit Program Classification System: The Taxonomy We've Been Waiting for , Introducing the Nonprofit Program Classification System: The Taxonomy We've Been Waiting for. *Nonprofit and Voluntary Sector Quarterly*, *30*(4), 781–793. doi:10.1177/0899764001304009

National Center for Charitable Statistics. (2006). *Guide to Using NCCS Data*. Urban Institute. Washington, DC.

Okten, C., & Weisbrod, B. A. (2000). Determinants of donations in private nonprofit markets. *Journal of Public Economics*, *75*(2), 255–272. doi:10.1016/S0047-2727(99)00066-3

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). doi:10.3115/v1/D14-1162

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. doi:10 . 1007 / BF00116251

Roeger, K. L., Blackwood, A. S., & Pettijohn, S. L. (2015). The Nonprofit Sector and Its Place in the National Economy. In J. S. Ott & L. A. Dicke (Eds.), *The Nature of the Nonprofit Sector* (Third edition, pp. 22–37). Boulder, CO: Westview Press.

Salamon, L. M. [Lester M.], & Anheier, H. K. (1992). In search of the non-profit sector II: The problem of classification. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, *3*(3), 267–309. doi:10.1007/BF01397460

Salamon, L. M. [Lester M], Anheir, H. K., & coaut. (1996). *The international classification of nonprofit organizations ICNPO-Revision 1, 1996*. Baltimore, Md: The Johns Hopkins University Institute for Policy Studies.

Sloan, M. F. (2009). The Effects of Nonprofit Accountability Ratings on Donor Behavior. *Nonprofit and Voluntary Sector Quarterly*, *38*(2), 220–236. doi:10.1177/0899764008316470

US Internal Revenue Service. (2013). IRS Static Files No. 2013-0005. https://www.irs.gov/pub/irs-wd/13-0005.pdf.

US Internal Revenue Service. (2014). Exempt Organizations Business Master File Information Sheet. https://www.irs.gov/pub/irs-soi/eo_info.pdf.

US Internal Revenue Service. (2018). 2017 Instructions for Form 990-EZ. https://www.irs.gov/pub/irs-pdf/i990ez.pdf.

Vakil, A. C. (1997). Confronting the classification problem: Toward a taxonomy of NGOs. *World Development*, *25*(12), 2057–2070. doi:10.1016/S0305-750X(97)00098-3

Zhang, Y., & Wallace, B. (2015). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. *arXiv:1510.03820 [cs]*. arXiv: 1510.03820 `[cs]`