Voluntas
https://doi.org/10.1007/s11266-021-00420-z

RESEARCH PAPERS

How to Code a Million Missions: Developing Bespoke Nonprofit Activity Codes Using Machine Learning Algorithms

Francisco J. Santamarina[1] · Jesse D. Lecy[2] · Eric Joseph van Holm[3]

Accepted: 17 September 2021

Abstract National Taxonomy of Exempt Entities (NTEE) codes have become the primary classifier of nonprofit missions since they were developed in the mid-1980s in response to growing demands for a taxonomy of nonprofit activities (Herman in Nonprofit and Voluntary Sector Quarterly 19(3):293–306, 1990, Barman in Social Science History 37:103–141, 2013). However, the increasingly complex nature of nonprofits means that NTEE codes may be outdated or lack specificity. As an alternative, scholars and practitioners can create a bespoke taxonomy for a specific purpose by hand-coding a training dataset and using machine learning classifiers to apply the codes to a large population. This paper presents a framework for determining training set sizes needed to scale custom taxonomies using machine learning algorithms.

Keywords Nonprofit organizations · Classification · Machine learning · Custom taxonomies

✉ Francisco J. Santamarina
fjsantam@uw.edu

Jesse D. Lecy
jdlecy@asu.edu

Eric Joseph van Holm
evanholm@uno.edu

[1] Evans School of Public Policy and Governance, University of Washington, 4105 George Washington Lane Northeast, Seattle, WA 98105, USA

[2] Watts College, Arizona State University, 411 N. Central Ave., Suite 750, Phoenix, AZ 85004-2163, USA

[3] Department of Political Science, Urban Entrepreneurship and Policy Institute, The University of New Orleans, 256 Milneburg Hall, New Orleans, LA 70148, USA

Introduction

Current taxonomies used to categorize nonprofit missions are useful but limited in their ability to capture important dimensions of nonprofit activities. The development of meaningful new taxonomies in the US context has been hindered by the daunting task of hand-coding mission statements for close to 1.5 million active nonprofits. The IRS release of machine-readable datasets that contain text fields of mission statements and program service accomplishments has created opportunities to leverage natural language processing and machine learning techniques to automate classification. Specifically, scholars can develop new taxonomies using a small sample of organizations, then use machine learning algorithms to scale classification by using the hand-coded sample as a training dataset. To facilitate the responsible dissemination of these tools, we present a guide for working with nonprofit text as data, a framework for understanding accuracy of classification as a function of training dataset size, and benchmarks to guide data requirements and performance expectations for projects developing new activity codes using machine learning techniques.

Most nonprofit studies that group organizations by activity or subsector currently use the National Taxonomy of Exempt Entities (NTEE) classification schema, the IRS Tax Exemption codes, or the International Classification of NonProfit Organizations codes (ICNPO). However, each has significant limitations. NTEE codes were developed by a group of scholars, funders, and nonprofit professionals in 1987 and are assigned to nonprofit organizations by IRS staff when they receive 501(c) tax-exempt status (Barman, 2013; Herman, 1990). The NTEE classification system consists of 26 major purpose codes (codes A–Z) and sub-classification within each, resulting in hundreds of unique

Published online: 07 October 2021

codes. As such, they are often aggregated into 10 or 12 broad nonprofit subsectors. Each nonprofit is assigned a single NTEE code, which tends to over-simplify their dynamic and evolving missions.

Whereas NTEE codes are assigned to a nonprofit based upon their described activities, the IRS Exempt Purpose Code is the self-reported qualification for tax-exempt status. When filing IRS Form 1023 founders report whether their organizational purpose fits into one or more of the eight conditions defined by the US tax code: religious, educational, scientific, literary, public safety, sports, or preventing cruelty. These exempt purposes codes represent the legal justification for tax-exempt status, so organizations must supply documentation of their activities and purportedly do not have an incentive to misdirect. Whereas the NTEE classification was developed by social scientists drawing upon measurement theory and instrument design the Exempt Purpose Codes were developed by tax professionals for pragmatic and not scholarly purposes.

Finally, the International Classification of Nonprofit Organizations (ICNPO) was created at a similar time as the NTEE system, but whereas NTEE codes were developed for administrative purposes within the context of a single country and thus reflect the peculiarities of law and society in the US, the ICNPO was designed with the global voluntary or independent sector in mind and was intended for cross-national comparative work (Salamon & Anheier, 1996).

Third sector scholarship has benefitted in immeasurable ways from the ability to group organizations into industries or subsectors. However, classification will always be an imperfect science. A classification system might be theoretically under-specified in a way that does not capture the most salient dimensions of nonprofit activity. Survey instruments used to assign organizations to categories can be confusing or lack specificity, making it difficult to apply consistently in real world contexts. More importantly, the sector has changed in the 35 years since the taxonomies were introduced and missions of specific nonprofits have evolved.

Some of the major challenges of classification by humans include validity and reliability. A taxonomy has validity if it is a meaningful way to organize the world. Construct validity can be established in a variety of ways with predictive validity being one of the most common approaches in social science. Does knowing that a person belongs to a specific group in a classification system predict behavior or provide objective information about them? For example, art nonprofits and human service nonprofits tend to have distinct revenue models. Knowing their type reveals useful information.

Reliability describes whether that construct can be measured consistently with a survey instrument or protocol. Important reliability considerations include:

- Inter-coder reliability: how consistently do different individuals categorize cases when using an instrument to classify a set of observations?
- Inter-temporal reliability: how stable is a latent construct over time?
- Latent versus hidden constructs: latent constructs are meaningful traits that are not directly measured but can be inferred from other variables or specialized instruments (e.g., IQ questionnaires to measure deductive reasoning). Hidden constructs are also meaningful traits but they cannot be observed. In this context, activity codes are latent if they are not directly disclosed by the nonprofit but observable using mission text. Organizational traits that are not *reliably* observable from mission statements are referred to as hidden classes, suggesting not that the nonprofit is withholding information or that they are not observable with other data, but that the classification task is not feasible with the current mission statement data.

New classifications have been introduced in efforts to improve upon the validity and reliability of dominant taxonomies or fine-tune specificity of in a specific research context. They will often leverage data sources and computational techniques new to the field. For example, there has been a rise in the computational classification of nonprofit mission statements (Fyall et al., 2018; Lecy et al., 2019a, 2019b; Ma, 2021). Though there are some concerns with the reliability of such methods, we found machine learning accuracy to be roughly comparable to inter-coder reliability results when looking at IRS tax-exempt purpose codes (Table 1). Specifically, humans and machines perform well with the same codes and struggle with the same codes. Human performance is a more realistic benchmark for machine learning approaches rather than perfect performance.

In this paper, we seek to move beyond classifications of nonprofit activities using generic and static mission codes and promote the development of bespoke taxonomies that better capture multidimensional and identity-oriented classes of activity. Specifically, we propose methods to assess the accuracy of machine learning models that automate the classification process in large databases. Can they be improved by pre-processing of text data using custom nonprofit lexicon files? How much nonprofit data is needed to be able to effectively train machine learning algorithms? And can we leverage these methods to develop new taxonomies that support a richer, multi-dimensional approach to the study of nonprofit mission identities?

Table 1 Comparison of machine learning and human accuracy

Schema	Code	ML accuracy	ICR	Difference
Tax-exempt purpose	Charity	0.84	0.79	0.05
Tax-exempt purpose	Religious	0.92	0.97	− 0.05
Tax-exempt purpose	Education	0.75	0.81	− 0.06
Tax-exempt purpose	Scientific	0.93	0.99	− 0.06
Tax-exempt purpose	Literary	0.96	0.98	− 0.02
Tax-exempt purpose	Safety	0.99	1.00	− 0.01
Tax-exempt purpose	Sports	0.96	0.96	0.00
Tax-exempt purpose	Cruelty	0.96	0.98	− 0.02
Serves vulnerable populations		–	0.87	
Sample size		3446	100	

As reported in Lecy et al. 2019b. The three authors separately coded the same random sample of 100 nonprofit mission statements, and reliability is measured using percentage of agreement. The implemented machine learning algorithm is Naïve Bayes. The final schema is a unique one, which the algorithm was not trained on and was used to further test ICR

See Appendix. In summary, accuracy reports the ratio of the sum of true positive and true negative cases divided by the sum of all positive and negative cases, both true and false

Current machine learning algorithms vary in the size of the training set required to calibrate the models. In addition to testing its accuracy, sensitivity, and specificity in applying the new taxonomy, we also explore how these standard metrics for machine learning algorithms change when the size of the training set increases, to identify zones of diminishing return. Preliminary analysis is done using naïve Bayes models, chosen due to their simplicity and high performance (Hand & Yu, 2001).

We expect to see the performance of the machine learning model will increase as the training size increases, with diminishing returns at thresholds determined by complexity and type of data and varying by type of algorithm used. We also expect that performing standard text cleaning steps on the data will also increase model performance. Initial results indicate that the machine learning model is almost as reliable as humans in coding for tax-exempt purposes. However, humans are better able to code missions in a way that balances sensitivity and specificity as considerations of accuracy.

Our work contributes standards that can be used to assess the robustness of new mission codes or taxonomies. Specifically, this paper can be used as a guide for determining the size of training datasets needed to achieve reliable automation of mission classification in large datasets. In addition, the analysis provides greater evidence of the utility of machine learning techniques for the study of nonprofits, and the relatively small amounts of data necessary to accurately develop novel classifications. By comparing performance on the unidimensional NTEE codes against the identity-oriented approach of non-exclusive tax-exempt purpose codes, we demonstrate considerations for a much richer analysis of the nonprofit sector that will yield new insights into activities and impact.

Study Methodology

The purpose of the study is to develop a framework that can be used to automate the coding of nonprofit activities described in mission statements or program activity fields using machine learning techniques. Specifically, we analyze the amount of data needed to effectively train a machine learning algorithm, defining effectiveness as accuracy of classification.

Existing NTEE and tax-exempt purpose codes are used to benchmark algorithmic performance as a function of data preparation steps and training dataset size.

Data

Our data consists of approved 1023-EZ filings for 2018 and 2019.[1] The IRS has publicly released meta-data files that include nonprofit names, mission statements, tax-exempt purpose codes, and NTEE codes. As described earlier, nonprofit founders self-report their purpose and NTEE codes. We omitted any filings that were missing mission statements and dropped duplicates (nonprofits can resubmit the form to provide updated information to the IRS). The study dataset consists of 104,072 newly formed nonprofits.

Classification is done using two text fields that represents information that is readily available for most nonprofits—their name and mission statement. Nonprofit names had 29.5 characters on average and mission statements 176.4 characters.

To test the efficiency of the models, we created three datasets: one with basic cleaning applied (“basic”), one cleaned following standard text analysis pre-processing steps (“standard”), and one using lexicon files specifically created for nonprofit missions (“custom;” see Paxton et al., 2019a, 2019b). For “basic,” the only adjustment was to make all source text lower case, then reduce sparsity (or cells in the data with a count of 0) by removing features that appear less than 100 times and in less than 100 documents (both steps performed on all three datasets); no words were removed or character strings modified beyond these two steps. Table 2 presents a side-by-side comparison

[1] Form 1023-EZ meta-data was downloaded from the IRS website: https://www.irs.gov/charities-non-profits/exempt-organizations-form-1023ez-approvals

Table 2 Comparison of text pre-processing steps by dataset

Text pre-processing steps	Convert to lowercase	Sparsity reduction	Remove white space	Remove punctuation	Common phrases converted to n-grams	Word stemming (quanteda vs. Paxton)	Custom spell-check (Paxton)
Basic	X	X					
Standard (quanteda)	X	X	X	X	X	Q	
Custom (Paxton)	X	X	X	X		P	P

of cleaning steps. We consider features as elements that we will use to train the classifier algorithm; in this case, our features are types, "the class of all tokens containing the same character sequence" (Manning et al., 2009, p. 22). A token is a string of text that has been parsed into a meaningful unit. After cleaning, we converted the dataset into a document frequency matrix and merged it with the original dataset of approved filings.

The data in "standard" and "custom" were cleaned using the R package quanteda (Benoit et al., 2018). For "standard," in addition to converting all source text to lower case, we applied common text pre-processing steps, including removing unnecessary white space and characters that consist of punctuation, numbers, symbols, and separators, while concatenating characters separate by a hyphen (e.g., "self-aware" becomes "selfaware"). Non-breaking spaces and common stopwords in the English language were also removed, using quanteda's dictionary (derived from Lewis et al., 2004). We reviewed the 100 *n*-grams (or "short subsequences of characters"; Manning et al., 2009, p.26, defining character *k*-grams) of $n = 3$, or 3-g, with the highest counts or appearances in the data, and identified a noticeable boundary at 150 counts. We then reviewed 3-g that had frequencies of at least 150 to determine if they made sense to be treated as one token. This process was repeated with 2-g that had counts of at least 600, another noticeable boundary. The relevant character sequences were rewritten as a single token for the appropriate 3-g and 2-g, in that order. Any spaces remaining within tokens were then removed, and tokens were stemmed using the default quanteda stemming tool. Stemming is an attempt to derive the roots or common character sequences of words by removing trailing characters that denote distinctions irrelevant for our study; for example, "profess", "professing, and "professes" would thus become "profess", whereas "professor" and "professors" would become "professor."

Steps for "custom" deviated for those from "standard" in two ways. Before removing non-letter characters, Paxton et al. (2019a) mission glossary was used to correct spelling errors and perform other corrections to the text. Instead of looking for the most common 3-g and 2-g to condense into single tokens and applying the default quanteda stemmer, we applied Paxton et al. (2019b) mission stemmer.

All three datasets then went through the same, final pre-processing steps. The tax-exempt purpose codes consist of eight non-mutually exclusive binary categories (a nonprofit can meet more than one requirement for 501c status). The NTEE codes were compiled into the 10 NTEE major groups using the crosswalk provided by the NCCS (Jones, 2019) and converted to 10 new binary variables indicating if the nonprofit belonged to a given category (1 = yes, 0 = no). The original 1023-EZ dataset also included variables as to whether an organization is classified as a public charity or private foundation[2] ("onethirdsupportpublic", "onethirdsupportgifts") and the activities it has or plans on engaging in[3] ("disasterreliefyes", "donatefundsyes"). These 22 variables were merged with the corpus as columns of binary variables prior to applying analytical techniques.

Techniques

We examine the relationship between the size of a training dataset and classifier accuracy using a bootstrapping approach to measure average accuracy and incrementally increasing the training dataset size. A training dataset of size N was randomly sampled from the full set of 104,072 documents in each corpus without replacement. We applied the default Naïve Bayes classifier algorithm in the quanteda package to predict the 22 categories of interest, or prediction classes, from the features extracted from the mission text (Benoit et al., 2018). Accuracy was determined using an independent testing dataset of 20,000 documents for all iterations. The training dataset started with 4000 processed mission statements, and increased by increments of 4000 up to a maximum of 80,000 mission statements in

[2] See "Part IV. Foundation Classification" in the instructions for the 1023-EZ form (IRS, 2018).

[3] See the entries for line 7 and line 12 in "Part III. Foundation Classification" (IRS, 2018).

the training set. Sampling was repeated 100 times in a bootstrap approach to create a distribution of accuracy scores associated with each training dataset size. The procedure was repeated for the binary variables used to represent each of the tax-exempt purpose codes, NTEE major group codes, and the several additional organizational variables from the 1023-EZ meta-data. We ran the bootstrapped classifier on the simulation cluster of terminal servers offered by the University of Washington's Center for Studies in Demography and Ecology. The bootstrapped classifier was applied in parallel to the three datasets using the R packages snow (Tierney et al., 2018) and parallel (R Core Team, 2020) on 66 cores, and each core was dedicated to predicting one of the 22 classes for each of the three datasets.

Metrics regarding classifier performance were generated using the R package caret (Kuhn, 2008). We evaluated the predictive performance of the algorithm for a given code based on balanced accuracy, a metric which considers Type I and Type II errors (see the Appendix). Conventional accuracy measures for classification algorithms can inflate performance when using an imbalanced dataset, one in which there is not an equal number of observations from the classes in question in the training dataset (Brodersen et al., 2010). Because we randomly assign cases to the training dataset such that at least one observation of the code in question (ex. NTEE, tax-exempt purpose) is present in both the training and testing dataset for each iteration, it seems reasonably probable that there will be at least one instance of the training dataset being imbalanced. As a result, using the traditional accuracy measure will result in a biased measure favoring whichever class is more frequent in our dataset (Brodersen et al., 2010), and so the balanced accuracy measure provides a more valid measure of the algorithm's performance.

We analyzed the performance with a focus on identifying thresholds of marginal returns or gains for improving the algorithm's predictive power. We created visualizations and the figures in this article using the base R visualization libraries in addition to the packages scales (Wickham & Seidel, 2020) and ggplot2 (Wickham, 2016).

Results: The Ability to Predict Codes

By applying the techniques above, we compiled 1,848,000 observations capturing data related to:

- 3 datasets
- 22 prediction classes
- 20 increments in training dataset size (from 4000 to 80,000 processed mission statements, in unit increases of 4000)
- 100 bootstraps at each training dataset size
- 14 different metrics of classifier performance

We then subset the data to only include data for the balanced accuracy metric and identify means, minima, maxima, and standard deviations for each distribution of bootstraps, resulting in a summary dataset of 1320 observations. Figure 1 visualizes the range of averaged balanced accuracy values, where boxes summarize the range of averaged values for prediction classes and boxes are separate by training dataset size (horizontally) and dataset (vertically).

When comparing averaged balanced accuracy results for the 22 prediction classes, we see that, overall, gains in the mean of the performance metric for a given code can vary widely. As training dataset size increases in units of 4000 observations, the range of averaged balance accuracy values seems to increase, though the interquartile range and median also shift higher. The interquartile range's upward movement slows down toward the middle of the range of training dataset sizes, suggesting a gradual plateauing in improvement as the size increases.

Training datasets are created by humans manually coding data and are thus time-intensive and a nontrivial labor input, especially if codes are cross-validated by multiple humans. As a result, the goal of this study is to identify the minimum number of cases needed to achieve a reasonable threshold of accuracy. Although more training data are always better, there is a rapidly diminishing return to new additional units of training data. We can borrow insights from the mathematics of profit maximization in micro-economics[4] to identify a criterion for optimal training dataset size. We find that returns to additional units of training data diminish rapidly once the new unit of training data fails to improve accuracy by at least 0.5 points (100 being perfect accuracy). The last unit of training data that achieves at least 0.5 points of accuracy is considered the optimal size. By comparing optimal training dataset size across 22 different activity codes and three data preprocessing approaches (basic, standard and custom), we can produce a reasonable benchmark or rule of thumb for training dataset size.

There is a possibility of too little salient information or too much noise in a training dataset creating a hidden construct scenario, resulting in poor classifier performance. In these instances, model performance does not improve significantly with additional training data, so optimal training dataset size will be low because the return on investment of additional data is so low. As a result, these

[4] Profits are maximized when firm production levels reach the point of diminishing returns to labor or capital, which is determined by identifying the point that the second derivative of the production function is equal to zero.

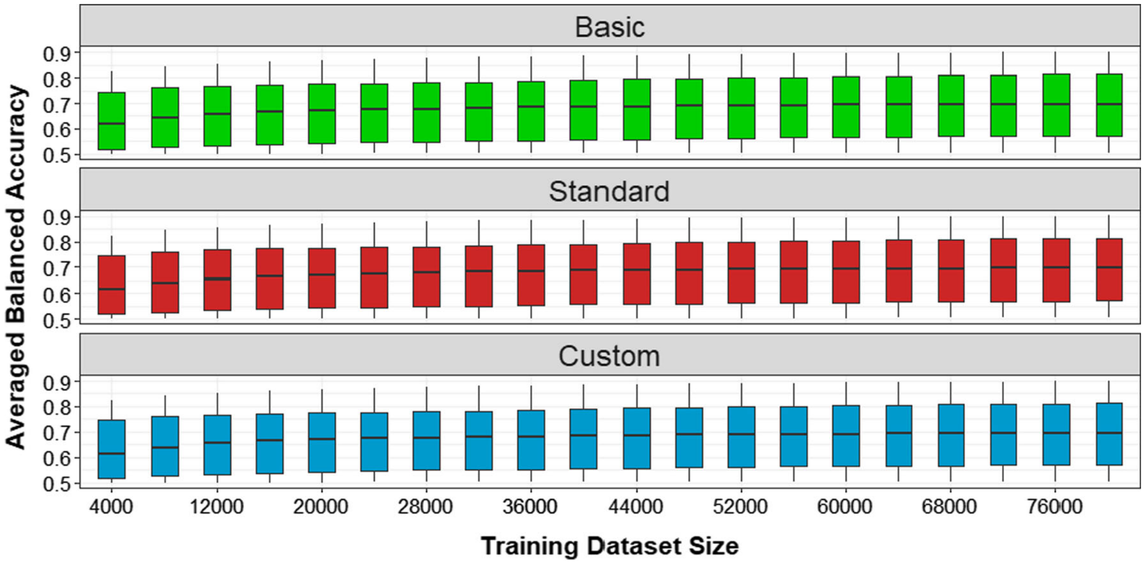

Fig. 1 Average balanced accuracy for each training dataset size. *X*-axis values report the training dataset size. The *y*-axis values report the balanced accuracy value. The box plots represent the range of balanced accuracy scores for each size, averaged by dataset and by the 22 total NTEE major groups, tax-exempt purpose codes, and 1023-EZ variables

cases can deflate estimates of appropriate dataset size and bias the benchmark downward. To compensate, we impose a threshold of 60% balanced accuracy and treat models with lower performance as cases of hidden constructs, where the input data is insufficient for the classification task. Our assumption is that the remaining cases above the 60% threshold describe dimensions of nonprofit activity that are communicated in the mission statements, and thus performance will be above 60%, i.e., greater than a purely random chance. Researchers developing new activity codes would select appropriate text that captures salient dimensions of mission, and thus the set of classes above the threshold is appropriate for benchmarking purposes.

Our proposed optimal training dataset sizes are presented in Fig. 2. Of the nine hidden classes, three are NTEE Major 10 codes (international, mutual benefit, and unknown), two are tax-exempt purpose codes (public safety and disaster relief), and the rest are the four 1023-EZ-specific categories. Among the visible codes, there are seven NTEE major 10 codes and 6 tax-exempt codes. The highest average balanced accuracy at a given 0.5 threshold size is 86.57%, and the lowest for a visible class is 60.75%. The highest for a hidden class is 56.07%, and the lowest overall is 50.11%.

The average optimal training dataset size is around 17,000 mission statements, with some slight variation between the three data preparation and cleaning methods. There were six prediction classes for which the three data preparation methods yielded similar results, four of which are tax-exempt purpose codes. Standard cleaning reached the threshold at the smallest size twice, custom once, and basic three times. Standard did so by the largest margin, an average of 6000 fewer cases than the nearest cleaning method.

Two cases stand out in sharp juxtaposition. For Religious, standard cleaning was the best performer of the three methods, and the algorithm reached the threshold roughly twice as fast (or with half as much training data) as with the other two. This is probably due to the use of *n*-grams to capture and denote common concepts consisting of multiple terms, especially ones used frequently in mission statements of Religious nonprofits. For Scientific, custom cleaning underperformed basic by ~ 12,000 observations and standard by ~ 8000 observations. This discrepancy is probably due to some combination of the custom stemmer and glossary in conjunction with vocabulary idiosyncratic to the scientific community. The overall low balanced accuracy of the model suggest that the text is borderline insufficient to produce reliable, non-random predictions.

Visible NTEE Major 10 classes varied broadly in terms of which cleaning method yielded the earliest threshold arrival: the three methods tied twice out of seven classes, basic was best twice, and standard and custom were each best once but tied once in out-performing basic. For the visible tax-exempt purpose classes, the three methods tied for four out of 6 classes. The extreme cases of Religious and Scientific were the other two classes.

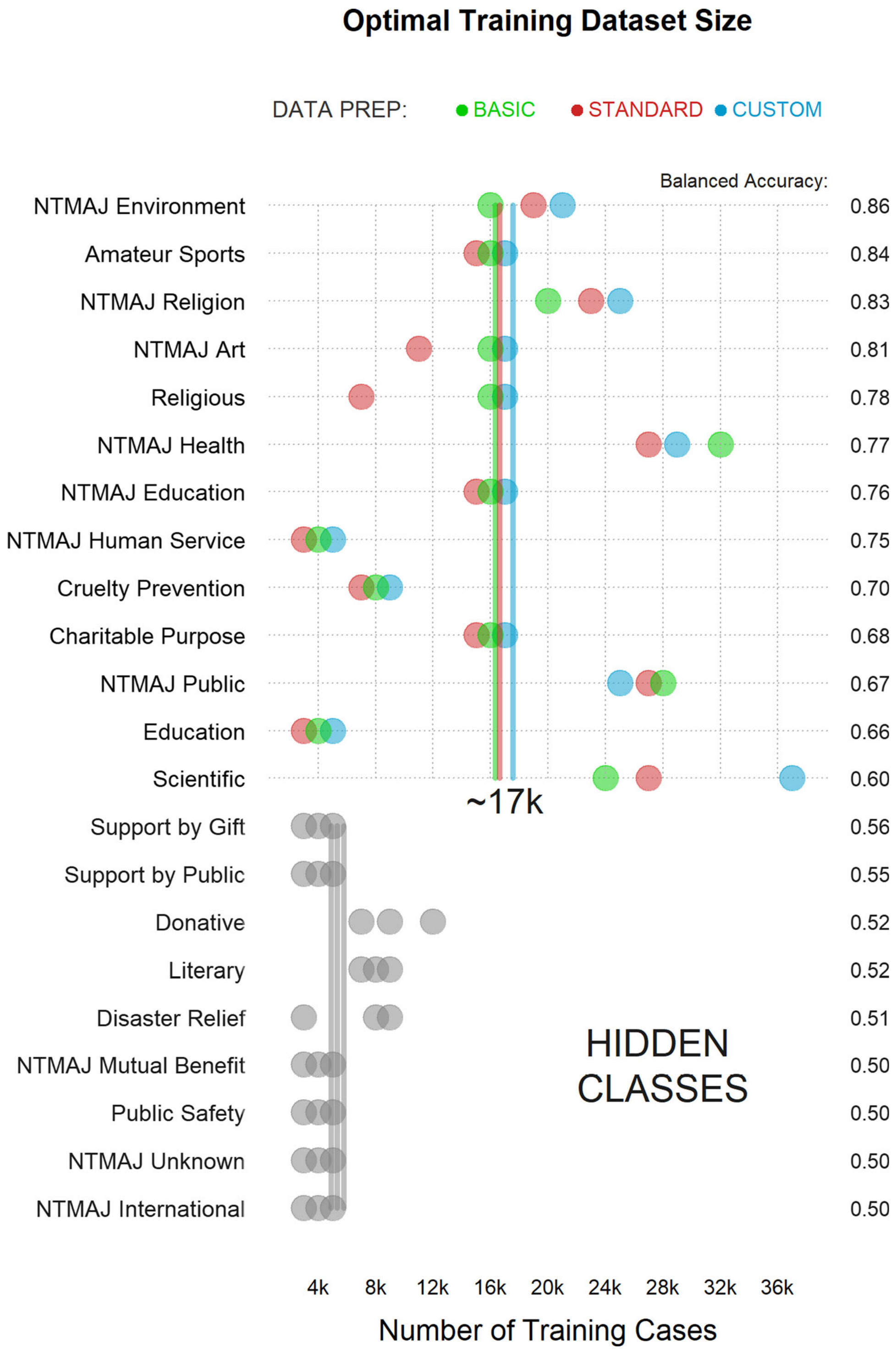

Fig. 2 Optimal training dataset sizes for prediction classes, as determined by the point at which increasing the training dataset size yields less than a 0.5 percentage point increase in balanced accuracy. The accuracy metric on the right reflects model performance associated with the optimal training dataset size, not the maximum accuracy possible with larger training datasets. Vertical lines indicate the average optimal training dataset size by data preparation step (corpora dataset). Hidden classes are those where the optimal training dataset size occurs at an average balanced accuracy of less than 60%. Please note that data were partially jittered to improve visual comprehension. See tutorial for further justifications and explanations

Discussion

In this paper, we have proposed a framework for developing more flexible taxonomies using an identity-oriented conceptualization of nonprofit activities. The use of big data and machine learning algorithms, as demonstrated here, opens new avenues for research into nonprofits based on the intention and scope of their mission. Rather than treating NTEE codes as a monolith, future research can work to disaggregate organizations based on patterns of activity, communities they serve, or other features of their stated missions or program descriptions. This research establishes a reasonable benchmark for the amount of hand-coded training data needed for such endeavors.

Nonprofit identities are not constrained to a single dimension of activity. More expressive taxonomies reflect the reality that organizations, like people, have many dimensions to their personalities. New taxonomies can be evaluated using information-theory criterion (whether information is gained by separating or combining categories), alternating between categorical (mutually exclusive levels) and multiple binary (not mutually exclusive levels) dimensions where appropriate, and addressing issues with dimensionality reduction where appropriate.

As seen above, cleaning methods vary less than expected in affecting algorithm performance increases. It is important to note that the three cleaning methods were not compared against a baseline of raw text. In other words, any cleaning seems to be useful, at minimum in making the data importable to algorithms. The unidimensional NTEE codes' variation in benefits from cleaning methods suggest the need to situate cleaning within the context of the prediction class, e.g., identifying *n*-grams for specific codes rather than across the full corpus. The non-exclusive tax-exempt purpose codes mostly showed minimal difference in cleaning method, yet they also showcased the extremes of optimal and perhaps inappropriate applications. Consolidating 2-g and 3-g as part of the standard methods may have also contributed to improvements, as this reduces conceptual sparsity by indicating that multiple character sequences capture a shared concept. We expect that consolidating more *n*-grams would increase performance gains, and that using the custom cleaning methods in addition to consolidating *n*-grams would yield the greatest improvements. In general, the data pre-processing steps were helpful but produced modest gains in accuracy relative to increases in training dataset size.

Using the set of 22 pre-existing activity codes and a resampling process that varies the training dataset size, we have identified a benchmark regarding the training dataset size for applying a custom taxonomy to nonprofits using machine learning approaches. All else equal, more training data is always better. But based on the bootstrapping procedure, we determine that a rule of thumb of approximately 17,000 mission statements balances algorithmic accuracy with the labor cost of hand-coding training data. As Fig. 2 also shows, there is some variance in optimal training dataset size across classification processes, so the proposed benchmark does not guarantee economical use of data in all contexts, but 17,000 cases is a reasonable rule of thumb for planning purposes.

Specificity in language of mission statements may affect the marginal return from increasing the training dataset size. In particular, sectors or organizational populations where the code can be widely applied or experience a large diversity in mission statements and purpose codes may experience the least benefit from minimal dataset size increases: in other words, more data may always be better. As with using 60% balanced accuracy as a cutoff, scholars should ensure that the relationship between the mission statement text and the prediction class is strong enough to make reliable predictions.

Given the high correlation between machine accuracy and human ICR reported in Table 1, it is likely that classification challenges encountered by humans will also impact machine learning algorithms. Somewhat counterintuitively, the optimal training dataset size for models with lower performance (i.e., more noise due to low observability of the latent class in the mission text), such as those for hidden classes, are actually smaller because the point of diminishing returns is reached faster.

Reflecting on a previous study and the results here, we note that studies in our field that implement machine learning algorithms should thus consider appropriate stopping points to increase training dataset size: at a certain peak in marginal return, at a percentage for the target metric (ex. 95% balanced accuracy), or at a plateau in percentage gains for the target metric (ex. increasing sample size yields one percentage point increase).

The validity and reliability of custom taxonomies becomes important when trying to generalize results and determine whether they can accurately shed light on the greater population being studied. Our findings can help inform appropriate strategies to ensure that taxonomies leveraging the new sets of "big" data available to nonprofit researchers are sound in their validity, reliability, and testing. We do note the contextual limitations on our benchmark and that it was derived from data in English developed in the U.S., where the use of tax laws to define nonprofits is arguably idiosyncratic and has directly influenced the data used in this study. More work is needed to reproduce this approach using mission data in other contexts, such as geography (ex. Europe, China), language (ex. French, German), and taxonomical systems (ex. United Nations' International Standard Industrial Classification, International Classification of Nonprofit Organizations, North American Industrial Classification Codes).

More fundamentally, mission statements are dynamic—they can change over time, a fact to be aware of when using mission statements from longitudinal data consisting of the same organizations across time, such as the IRS 990-derived datasets. In contrast, NTEE and tax-exempt purpose codes are static. User input error and static versus dynamic missions represent two serious research challenges as more longitudinal nonprofit data becomes available, underscoring the potential advantages of algorithmic approaches to recoding data at scale as nonprofit scholars become more adept at machine learning approaches.

Appendix: Model Fit Formulas

Metric	Formula
Sensitivity	$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = \frac{\mathrm{TP}}{\mathrm{P}}$
Specificity	$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = \frac{\mathrm{TN}}{\mathrm{N}}$
Precision	$\mathrm{PREC} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$
Recall	$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = \frac{\mathrm{TP}}{\mathrm{P}}$
*F*1	$F_1 = \frac{2 \cdot \mathrm{PREC} \cdot \mathrm{REC}}{\mathrm{PREC} + \mathrm{REC}}$
Accuracy	$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{P}+\mathrm{N}}$
Balanced accuracy	$\mathrm{BA} = \frac{(\mathrm{Sensitivity} + \mathrm{Specificity})}{2}$
	$= \left[\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\right) + \left(\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right)\right] * \frac{1}{2}$
Error	$\mathrm{ERR} = \frac{\mathrm{FP}+\mathrm{FN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}} = \frac{\mathrm{FP}+\mathrm{FN}}{\mathrm{P}+\mathrm{N}}$

Source: Balanced Accuracy's first formulation comes from Kuhn (2019). The second formulation comes from Brodersen et al. (2010). Other formulas come from Saito and Rehmsmeier (n.d.).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11266-021-00420-z.

Acknowledgements Special thanks to the ARNOVA 2020 Conference Doctoral Fellowship Program participants, ARNOVA 2019 Conference panel feedback, and the USC's Price School of Public Policy, The Center on Philanthropy & Public Policy's "Philanthropy & Social Impact: A Research Symposium" (March 15, 2019).

Funding Partial support for this research came from a Eunice Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant, P2C HD042828, to the Center for Studies in Demography & Ecology at the University of Washington.

Data Availability Data and code to replicate the results presented in the article are available via the authors' dedicated GitHub repository and Harvard Dataverse site: https://fjsantam.github.io/bespoke-npo-taxonomies/

Declarations

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

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*, 103–141. https://doi.org/10.2307/23361114

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software, 3*(30), 774. https://doi.org/10.21105/joss.00774

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition* (pp. 3121–3124). IEEE.

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.

Hand, D. J., & Yu, K. (2001). Idiot's Bayes—not so stupid after all? *International Statistical Review, 69*(3), 385–398.

Herman, R. D. (1990). Methodological issues in studying the effectiveness of nongovernmental and nonprofit organizations. *Nonprofit and Voluntary Sector Quarterly, 19*(3), 293–306. https://doi.org/10.1177/089976409001900309

Internal Revenue Service. (2018). *Instructions for form 1023-EZ: Streamlined application for recognition of exemption under section 501(c)(3) of the internal revenue code* (Cat. No. 66268Y). Retrieved from https://www.irs.gov/pub/irs-pdf/i1023ez.pdf.

Jones, D. (2019). IRS activity codes. Published January 22, 2019. https://nccs.urban.org/publication/irs-activity-codes.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*(5), 1–26.

Kuhn, M. (2019). The 'caret' package. "17 Measuring Performance." https://topepo.github.io/caret/measuring-performance.html.

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

Lecy, J. D., Santamarina, F. J., & van Holm, E. J. (2019b). The political economy of nonprofit entrepreneurship: Using open data to explore geographic and demographic dimensions of nonprofit mission [Paper presentation]. USC CPPP Symposium, Los Angeles, California.

Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research, 5*(Apr), 361–397.

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.

Manning, C. D., Schütze, H., & Raghavan, P. (2009). *Introduction to information retrieval*. Cambridge university press. Online edition. https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf.

Paxton, P., Velasco, K., & Ressler, R. (2019a). Form 990 Mission Glossary v.1. [Computer file]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor].

Paxton, P., Velasco, K., & Ressler, R. (2019b). Form 990 Mission Stemmer v.1. [Computer file]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor].

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Saito, T., & Rehmsmeier, M. (n.d.). Basic evaluation measures from the confusion matrix. https://classeval.wordpress.com/introduction/basic-evaluation-measures/.

Salamon, L. M. & Anheier, H. K. (1996). The International classification of nonprofit organizations: ICNPO-Revision 1, 1996. Working Papers of the Johns Hopkins Comparative Nonprofit Sector Project, no. 19. Baltimore: The Johns Hopkins Institute for Policy Studies.

Tierney, L., Rossini, A. J., Li, N., & Sevcikova, H. (2018). snow: Simple network of workstations. R package version 0.4–3. https://CRAN.R-project.org/package=snow.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Wickham, H., & Seidel, D. (2020). scales: Scale functions for visualization. R package version 1.1.1. https://CRAN.R-project.org/package=scales.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.