Julia Litofcenko and Dominik Karner and Florentine Maier

Methods for Classifying Nonprofit Organizations According to their Field of Activity: A Report on Semiautomated Methods Based on Text

Paper

Methods for Classifying Nonprofit Organizations According to their Field of Activity: A Report on Semiautomated Methods Based on Text

There are various methods for classifying nonprofit organizations (NPOs) according to their field of activity. We report our experiences with using two semi-automated methods based on textual data: rule-based classification, and machine-learning with curated keywords. We use those methods to classify Austrian nonprofit organizations based on the International Classification of Nonprofit Organizations (ICNPO). Those methods can provide a solution to the widespread research problem that quantitative data on the activities of NPOs are needed but not readily available from administrative data, long high-quality texts describing NPOs' activities are mostly unavailable, and human labor resources are limited. We find that in such a setting, rule-based classification performs about as well as manual human coding in terms of precision and sensitivity, while being much more labor-saving. Hence, we share our insights on how to efficiently implement such a rule-based approach. To address scholars with a background in data analytics as well as those without, we provide non-technical explanations and open-source sample code that is free to use and adapt.

Introduction

The increasing availability of large amounts of rich and growing administrative or otherwise process-generated data, often referred to as big data, has prompted scholars to consider new ways of using these data for research on nonprofit organizations (NPOs) and civil society (see, for example, Lecy & Thornton, 2016; McDonnell & Rutherford, 2018). One important piece of information concerns NPOs' fields of activity. Unfortunately, many available data sets do not contain such information in readily usable form, because classification by fields of activity is missing or of poor quality (see, for example, Grønbjerg & Paarlberg, 2002:588 on consistency problems with NTEE classifications in IRS data in the U.S.). The research task of complementing existing data sets of NPOs with an additional variable that indicates NPOs' main field of activity (or all their fields of activity, for more detailed analyses) is therefore common. However, there is yet no shared understanding of methods to accomplish this task.

This research note aims at contributing to a common understanding of computational methods for classifying NPOs according to their field of activity, based on textual data about those NPOs. Specifically, we discuss two approaches that represent the two main families of computational methods for classification (Zhai & Massung, 2016:300-302): so-called rule-based methods, and machine learning methods. As a rule-based approach, we discuss classification using a decision-tree algorithm that was generated by humans with background

information. As a machine learning approach, we discuss classification using keywords curated by humans and a decision-tree algorithm that was generated based on statistical properties.

We thereby focus on two semi-automated approaches that are useful in research settings where long high-quality texts about NPOs' activities – such as mission statements – are not available. We only use the NPOs' names as input data. These semi-automated approaches are a feasible alternative when fully automated approaches based on longer texts and machine learning are not possible (on such approaches see Lepere-Schloop, 2017; Lepere-Schloop, Zook, & Bawole, 2018).

As classification scheme for NPOs' field of activity we rely on the well-established International Classification of Nonprofit Organizations (ICNPO) as described by Salamon and Anheier (1992). The ICNPO was developed in the course of the Johns Hopkins Comparative Nonprofit Sector Project, which involves 45 countries. It has proven its applicability in a wide range of cross-country comparative studies, and in national accounting and statistics in line with recommendations by the OECD and UN. The latter recommendations include some modifications to the ICNPO, to more precisely cover cooperatives and similar market producers (United Nations, 2018:76-77).

We report the results of assigning NPOs to one single class at the level of ICNPO groups, except for the group „culture and recreation", where, due to the many organizations in this group we further discern between subgroups. Table 1 provides a visual overview of this classification system. The classification exercises reported in this paper could easily be adapted to include more or different subgroups, or to assign NPOs to several classes.

Table 1: ICNPO groups and subgroups used

(Sub-)group number	(Sub-)group name
1 000	Culture and recreation
1 100	Culture and arts
1 200	Sports
1 300	Other recreation and social clubs
2 000	Education and research
3 000	Health
4 000	Social services
5 000	Environment
6 000	Development and housing
7 000	Law, advocacy and politics
8 000	Philanthropic intermediaries and voluntarism promotion
9 000	International
10 000	Religion
11 000	Business and professional associations, unions
12 000	Not elsewhere classified

We develop our argument by first describing the particularities of our research setting, our sample, and the procedure for identifying organizations' true ICNPO category. Then, to establish a benchmark for the semi-automated approaches, we present performance metrics of manual human coding. Next, we explain the ideas behind rule-based classification and classification based on machine learning. We report on our experience of applying these approaches, and we compare their performance metrics, efficiency, and transparency. We find that, considering these three criteria, rule-based classification is the most sensible approach in our empirical setting. We conclude by discussing further strengths and limitations of the various approaches, and by providing recommendations for efficiently implementing a rule-based approach.

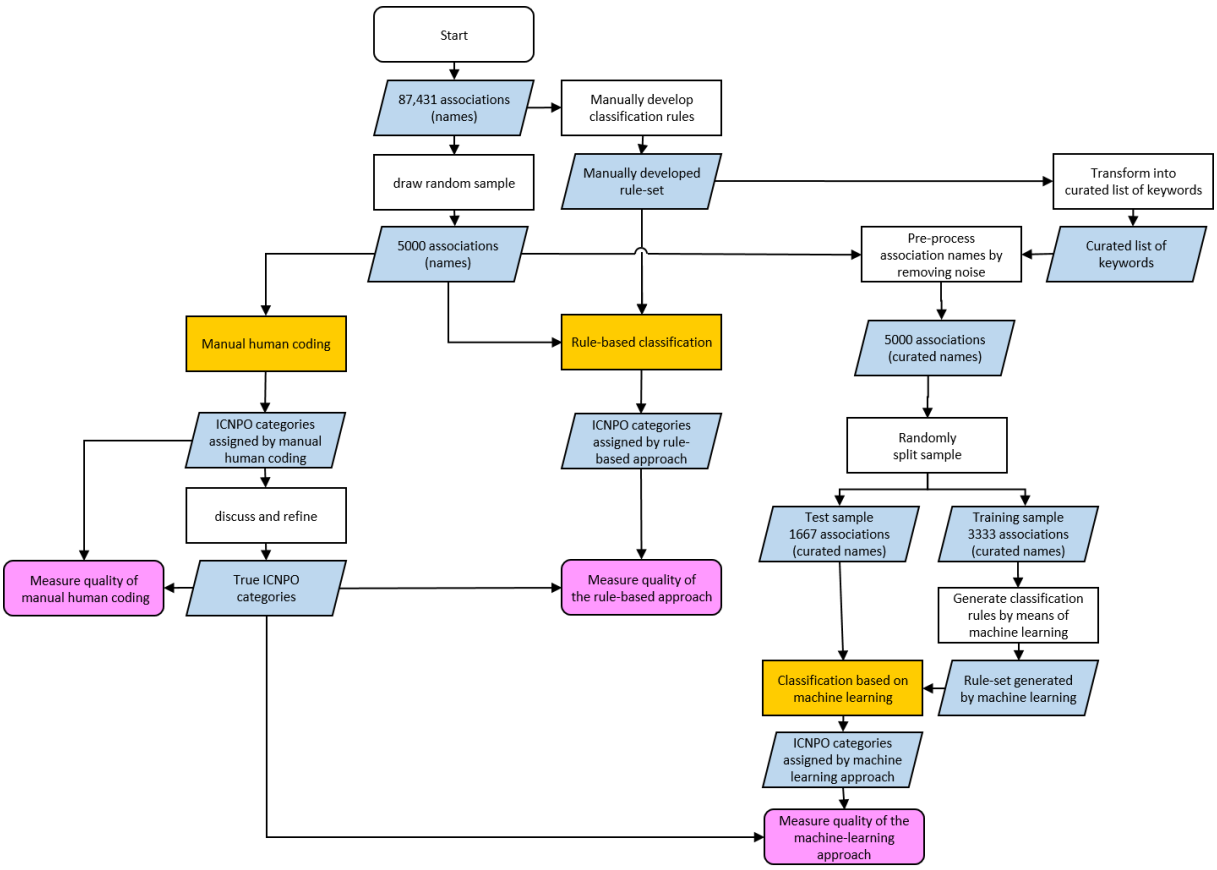
We expect our insights to be useful for various research scenarios: In studies investigating causal relationships, data on NPOs' activities provide an important control variable. Also, knowledge about NPOs' activities is often desirable for its own sake, e.g., for mapping purposes. In particular, the rule-based approach proposed here may be of interest to experts involved in the preparation of national accounts, who seek to identify and classify „non-profit and related institutions" following recommendations by the United Nations (2018). The classification scheme recommended by the UN is a revised version of the ICNPO that we use here, and is compatible with the ICNPO.

Setting, sample, and way of identifying true ICNPO categories as the point of reference

To put the ensuing discussion of methodological alternatives into context, we start by clarifying relevant aspects of our empirical setting and sample. We explain how we identified an organization's true ICNPO category, and how we thereby determined the point of reference for measuring the performance of various classification methods.

The empirical setting for our classification exercise is Austria. We focus on nonprofit associations, because 99% of Austria's NPOs have this legal form (Vandor, Traxler, Millner, & Meyer, 2017). There is a legally prescribed register of associations, which documents certain key data of all associations in Austria: the association's name, address, founding year, and legal representatives. The original register lies within the Ministry of the Interior and is not publicly available. However, via the business information publisher *Compass Verlag GmbH* we were able to obtain a database that almost completely mirrors the official register. We ensured the quality of the data by comparing it with publicly available Ministry figures on the total number of associations. We worked with data as of November 24, 2017, containing a total population of 87,431 active associations. From this population, we drew a random sample of 5,000 associations to measure the performance of the various classification approaches (see Figure 1).

Figure 1: Flow chart of the research process



To measure the performance of various classification approaches, we needed to establish a point of reference. This point of reference is the true ICNPO category of every NPO, so that it is possible to measure how many NPOs are classified correctly by an approach (see Figure 1 for an overview of the complete research process). Austrian administrative data does not include ICNPO categories, so we chose the following procedure: Each of the three authors independently classified each NPO. For doing so, we relied on the organization's name, if we recognized it and had additional background knowledge of the organization's activities. We also relied on the organization's name if it appeared informative enough in itself. When we were uncertain about an NPO's activities, we independently conducted desk research to clarify the issue. If all three coders unanimously assigned the same ICNPO category to an NPO, we adopted this category as the true ICNPO category. In all other cases, we determined the true ICNPO category by discussing it in the research team, and if necessary conducting further desk research.

For the following explanations, it is important to note that Austrian law prohibits organizations from carrying names that are so misleading about their nature as to cause harm to the public. Such law is standard in countries with a developed legal system. Moreover, Austrian law is particularly strict on the naming of associations. Associations are not only forbidden to use misleading names; they must moreover carry names that indicate their purpose. To ensure that our findings are relevant also in countries without such strict association law, we additionally coded a random sample of 1,000 nonprofit associations from Germany[1]. German law just prohibits dangerously misleading organization names, but does not require associations to carry names that indicate their purpose. Indeed, we found that the German sample included more organizations with names that consist only of an abbreviation, only of the name of a little-known founder or beneficiary, or of a neologism without clear meaning. Compared to the Austrian sample, the German sample contained 3 percentage points more organizations whose names did not provide clear information about their purpose. Hence, if the methods presented below are applied in countries with more liberal naming laws, a deterioration in performance of the magnitude of ca. 3 percentage points can be expected for all classification methods based on NPO names.

[1] Scraped from https://www.vereinsverzeichnis.eu, last accessed 07.08.2019.

Manual human coding

Based on the manual classification work that we had done to determine the true ICNPO categories, we were able to measure the performance of manual human coding and thereby set a benchmark for assessing the performance of the semi-automated approaches. Manual human coding, in this sense, means classification *before* the discussions in the research team to assign true ICNPO categories (see Figure 1).

Table 2: Performance of individual human coders

Coder	Correctly classified (n)	Correctly classified %
Coder A	3961	79%
Coder B	4118	82%
Coder C	4371	87%
Total	**5000**	**100%**

Table 3: Overall performance of manual human coding

ICNPO Group		True ICNPO (n)	True ICNPO %	Sensitivity of mode of human coders %	Precision of mode of human coders %
1100	Culture	994	20%	92%	94%
1200	Sports	1061	21%	92%	96%
1300	Other Recreation and Social Clubs	909	18%	87%	84%
2000	Education and Research	299	6%	86%	92%
3000	Health	94	2%	70%	80%
4000	Social Services	385	8%	82%	91%
5000	Environment	84	2%	71%	92%
6000	Development and Housing	404	8%	85%	82%
7000	Law, Advocacy and Politics	187	4%	71%	83%
8000	Philanthropic Intermediaries and Voluntarism Promotion	6	0%	50%	100%
9000	International	75	2%	87%	88%
10000	Religion	90	2%	66%	89%
11000	Business and Professional Associations, Unions	350	2%	81%	82%
12000	Not Elsewhere Classified	62	1%	13%	100%
Total		**5000**	**100%**	**85%**	

Note: Sensitivity = TP/(TP+FN); Precision=TP/(TP+FP). TP=*true positive*, FN=*false negative*, FP=*false positive*.

As shown in Table 2, individual coders classified 79% to 87% of associations in the sample correctly. Among the coders, the percentage of correctly classified NPOs was obviously positively related to the amount of experience in the field of NPO research. A common method to measure the performance of human coding is to use the mode of human coders, i.e., the category assigned by all or most of the coders. When measured this way, manual human coding correctly classified 85% of the organizations (see Table 3). It assigned 11% of the

organizations to a false ICNPO category, and 4% could not be classified at all, because every human coder suggested a different category.

In terms of efficiency, manual human coding is very time-consuming. It took us approximately 120 person-hours to manually classify 5,000 NPOs. This work resulted only in the classification of those NPOs, not in an algorithm that could also be used to classify the full population. Coders have to invest time to develop thorough classification rules to ensure a minimum degree of inter-coder consistency. This work can hardly be outsourced to untrained staff, because coders need to have substantive knowledge of the nonprofit sector.

The transparency of manual coding is low. Written coding instructions will be either highly ambiguous or extremely extensive (hence complicated, hence error-prone). There is no way for outsiders to reconstruct the reasoning that led a coder to assign a particular NPO to a particular category. If systematic classification errors become apparent in retrospect, it is very time-consuming to correct them.

Since manual human coding, if not done by proficient coders, is quite inaccurate, inefficient and hardly transparent, we endorse computational approaches. In the following sections, we will present two such approaches: rule-based classification, and a machine learning approach based on manually curated keywords. We will report on our results and insights gained from applying them.

Rule-based classification

Rule-based classification is semi-automated classification based on manually created IF-THEN rules. A simple example of such a rule is: IF the organization's name includes the word „fan club", THEN assign the organization to the ICNPO category „other recreation and social clubs". As suggested by Zhai and Massung (2016:301), rule-based classification is likely to work well if the following criteria are met:

(1) Categories are clearly defined.
(2) Categories can be relatively easily distinguished based on surface features in the text (e.g., particular words).
(3) Researchers have sufficient domain knowledge to suggest many effective rules.

All three criteria were fulfilled in our case, as we intended to classify NPOs in Austria according to ICNPO categories based on the organizations' names:

(1) The ICNPO provides clearly defined categories.

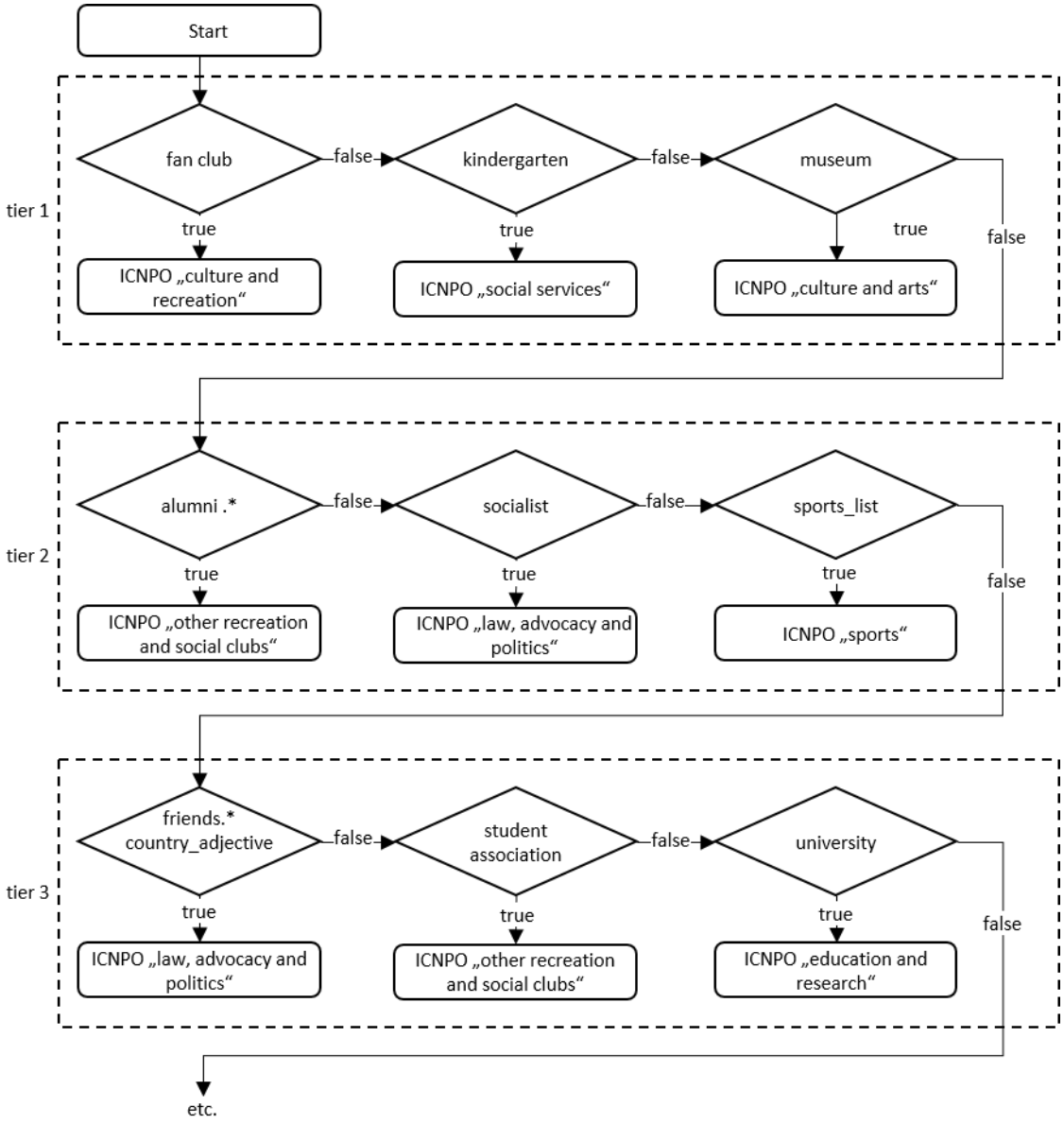
(2) Names of NPOs, in most cases, gave sufficient information to classify the organizations according to the ICNPO.
(3) The research team had background knowledge about the country's nonprofit sector, and had the possibility of doing additional desk research to clarify remaining ambiguities.

Researchers who wish to apply a rule-based classification approach in another setting will need to check whether these criteria are met. Moreover, for optimal results, classification rules must be established separately for different countries – or to put it more precisely, for each language region with a specific civil society tradition. These rules are based on texts that require a thorough understanding of the language and culture from which they originate.

The main work for implementing a rule-based approach is to develop a system of classification rules: Researchers manually look for suitable search terms and order them in appropriate tiers to build the rule-set. Figure 2 gives a simplified example of such a rule-set. It is a manually created decision tree with binary univariate splits at the nodes. Each decision node is an IF-THEN rule about a particular search term. IF a yet uncategorized organization has this search term in its name, THEN this organization goes to a specific ICNPO category associated with that search term. Hence, every node sorts out some cases. Then the next rule is applied to the remaining uncategorized organizations.

Rules within one tier are mutually exclusive. Hence, their order within the tier is not important. When rules are not mutually exclusive, i.e. when associations' names include two or more search terms, those search terms are ordered hierarchically in different tiers. For example, there are socialist student associations, whose work is mainly political. They are classified as belonging into the category of „law, advocacy and politics". Most other student associations are social clubs and therefore belong to the category of „other recreation and social clubs". Thus, the search term „socialist" needs to be placed in a higher tier than „student association".

Figure 2: Example of rule-based classification in the Austrian case



The rule-based algorithm was able to correctly predict the ICNPO category for 85% of the Austrian associations in the sample. Thus, rule-based classification produces results that are not inferior to manual human coding. Table 4 provides more detailed performance metrics. Most misclassifications occur in the category „not elsewhere classified“. This is because the algorithm assigns all unclassifiable organizations to this category. Hence, two kinds of organizations end up in this category: a large number of NPOs that actually belong to another category, and a small number of NPOs that also the human experts found to be truly unclassifiable.

Table 4: Performance of rule-based classification (column percent; figures are rounded)

		true ICNPO														
		1100	1200	1300	2000	3000	4000	5000	6000	7000	8000	9000	10000	11000	12000	% predicted ICNPO
predicted ICNPO	1100	90%	0%	0%	1%	1%	1%	0%	1%	6%	0%	1%	1%	0%	2%	19%
	1200	0%	90%	1%	0%	1%	0%	0%	0%	1%	33%	0%	0%	1%	2%	20%
	1300	1%	2%	86%	0%	2%	3%	4%	1%	3%	0%	1%	0%	2%	2%	17%
	2000	0%	0%	0%	86%	1%	1%	1%	0%	0%	0%	1%	2%	0%	2%	5%
	3000	0%	0%	0%	1%	85%	2%	2%	0%	0%	0%	1%	1%	1%	0%	2%
	4000	1%	0%	1%	1%	2%	85%	2%	1%	3%	17%	9%	1%	1%	2%	8%
	5000	0%	0%	0%	2%	0%	1%	79%	2%	1%	0%	0%	0%	1%	2%	2%
	6000	0%	0%	0%	0%	0%	0%	1%	80%	2%	0%	3%	0%	4%	0%	7%
	7000	0%	0%	1%	1%	0%	1%	0%	0%	64%	0%	5%	0%	1%	0%	3%
	8000	0%	0%	0%	0%	0%	0%	0%	0%	0%	17%	0%	0%	0%	0%	0%
	9000	0%	0%	0%	1%	0%	0%	0%	0%	2%	0%	53%	0%	2%	2%	1%
	10000	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	80%	1%	0%	2%
	11000	0%	0%	0%	1%	2%	0%	2%	2%	5%	0%	1%	1%	74%	2%	6%
	12000	6%	5%	10%	6%	5%	6%	8%	11%	13%	33%	23%	13%	12%	87%	9%
true ICNPO (n)		994	1061	909	299	94	385	84	404	187	6	75	90	350	62	5000
% true ICNPO		20%	21%	18%	6%	2%	8%	2%	8%	4%	0%	2%	2%	7%	1%	100%
precision		96%	98%	92%	94%	78%	85%	67%	93%	82%	100%	66%	90%	89%	12%	

Note: Sensitivity=TP/(TP+FN) in the diagonal, grey shadowed. Precision=TP/(TP+FP). TP=*true positive*, FN=*false negative*, FP=*false positive*.

Since, as mentioned above, Austrian law is particularly strict about the informative naming of associations, we cross-checked whether rule-based classification also works in a country with more liberal regulations. For this purpose, we applied the rule-set devised for Austria to the abovementioned random sample of 1,000 German associations. Before doing so, we had not only assessed the percentage of associations with names that would not satisfy legal requirements in Austria (3%). We had also determined the percentage of associations with names that exhibit differences in language use and culture compared to Austria (17%). For those associations, it would have been necessary to modify the rule-set, because although in both countries the same language is spoken, civil society landscapes differ considerably (Heitzmann & Simsa, 2004; Zimmer et al., 2004). With the unmodified rule-set we could correctly classify 64% of the German associations. This rate almost exactly equals the Austrian rate (85%) minus deductions for the poorly adjusted rule-set (17%) and the more liberal law (3%).

Classification based on machine learning and curated keywords

The term machine learning refers to a variety of methods for detecting patterns in large amounts of data. These methods apply statistical algorithms to find patterns in a so-called training sample and automatically formulate classification rules based on these patterns. These rules can be used to classify infinite amounts of further cases. Common machine learning methods used for text classification are Naïve-Bayes classifiers, decision trees, regression methods and neural networks (Lantz, 2015; Zhai & Massung, 2016).

As with any other statistical method, classification results based on machine learning will be satisfactory only when based on enough input data of good quality. In our setting we were faced with serious limitations of data, because we had no long high quality texts such as mission statements that contain information about NPOs' activities.

We experimented with various machine learning algorithms, using only organization names as input data, or using longer input texts obtained through web scraping[2] as input data. Results were unsatisfactory. Using a training sample of n=1,068 and a test sample of n=750[3], neither decision tree models nor naïve Bayes nor multinomial lasso regression-models classified more than 50% of the test sample correctly, neither based on organization names nor based on the longer texts obtained through web scraping.[4] Since other scholars had been able to achieve much better results for similar classification tasks with smaller sample sizes (Fisher, 2016; Lepere-Schloop, 2017), we concluded that not the sample size but the quality of input texts

---

[2] We obtained snippets from the search engine Bing through web scraping. Bing hosts an Application Programmer Interface (API) that allows using the search engine in an automated fashion. Our rationale for using those snippets was that the algorithms applied by large search engines are very good at summarizing relevant information from texts. The snippets that we thereby obtained contained on average 63 words per association. We prepared those texts using common techniques for text pre-preprocessing in bag-of-words models: stemming, and removing stop words, non-alphabetic characters and one-letter words (Kwartler, 2017; Lantz, 2015).

[3] These sample sizes are smaller than those reported in the rest of the paper because the search engine Bing did not find information about all of the associations in the complete sample.

[4] All machine learning models were based on bag-of-words representations of the text. The following pre-processing steps were applied: Removal of non-alphabetic characters, stop words; stemming; feature selection through tf-idf.

was the reason for the bad classification performance in our case. Pure machine learning approaches were not feasible in this setting.

We hence opted for a semi-automated machine-learning approach, which relies on manually curated keywords. In such an approach, the quality of the input texts is improved by reducing noise, i.e., removing all words that do not contain relevant information for assessing an NPO's field of activity. We implemented this approach by removing every word from the organizations' names that was not in the search term list developed for the rule-based approach (see Figure 1). Table 5 provides examples of how the relevant features of the organizations' names were pre-selected. A decision tree model[5] using the curated organization names as input performed far better than the abovementioned machine learning models. It correctly classified 77% of the organizations in the test sample (n=1,667). Table 6 shows detailed performance metrics. The performance of the model varies strongly across categories, and the variation does not seem to be driven by the number of cases in the respective categories.

Table 5: Examples of curated organization names

Original organization name	Curated association name
Studentensport	.*ensport.* .*sport.* .*student.*
GOLD - FINGER : gemeinnütziger Verein zur Förderung der Musikkultur in EUROPA	musikkultur musik.* .*kultur.* .*musi.*
Alumni der Akademie der bildenden Künste Wien	Alumni.* akademie künste.*
Bosniakische Kultur- und Glaubensgemeinschaft Oberland	glaubens.* bosniak.* kultur .

[5] C5.0 algorithm.

Table 6: Performance of decision tree classification with curated organization names (column percent; figures are rounded)

		true ICNPO														
		1100	1200	1300	2000	3000	4000	5000	6000	7000	8000	9000	10000	11000	12000	% predicted ICNPO
predicted ICNPO	1100	84%	0%	1%	1%	0%	1%	0%	0%	3%	0%	0%	11%	0%	0%	39%
	1200	1%	88%	2%	2%	3%	0%	0%	0%	3%	0%	0%	0%	2%	0%	20%
	1300	13%	11%	91%	10%	10%	14%	12%	31%	40%	100%	19%	17%	19%	100%	13%
	2000	0%	0%	1%	77%	7%	1%	6%	4%	1%	0%	0%	0%	1%	0%	5%
	3000	0%	0%	0%	0%	59%	2%	0%	0%	1%	0%	0%	0%	1%	0%	2%
	4000	1%	1%	1%	3%	14%	75%	3%	4%	1%	0%	0%	3%	3%	0%	7%
	5000	0%	0%	0%	1%	0%	3%	67%	0%	1%	0%	5%	0%	0%	0%	1%
	6000	0%	0%	1%	2%	0%	1%	9%	53%	4%	0%	0%	0%	5%	0%	4%
	7000	0%	0%	0%	2%	0%	2%	0%	1%	42%	0%	0%	6%	0%	0%	2%
	8000	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	9000	0%	0%	0%	1%	0%	0%	0%	0%	1%	0%	76%	0%	3%	0%	1%
	10000	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	57%	0%	0%	1%
	11000	1%	0%	2%	1%	7%	1%	3%	7%	3%	0%	0%	6%	66%	0%	6%
	12000	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
true ICNPO (n)		325	358	299	103	29	138	33	112	77	1	21	35	115	21	1667
% true ICNPO		19%	21%	18%	6%	2%	8%	2%	7%	5%	0%	1%	2%	7%	1%	100%
precision		95%	96%	54%	84%	77%	79%	73%	77%	76%	-	70%	91%	74%	-	

Note: Sensitivity=TP/(TP+FN) in the diagonal, grey shadowed. Precision=TP/(TP+FP). TP=*true positive*, FN=*false negative*, FP=*false positive.*

These experiences with machine learning models suggest that the quality of input texts is key, and that the quality can be improved substantially by pre-selecting the relevant features of the input texts. Nevertheless, the machine learning approach with curated organization names performed worse than the rule-based approach. This seems surprising at first glance, given that the machine learning algorithm determines the rules by going through a large number of possible combinations of search terms, and selects those with the highest predictive power for the outcome variable. However, our data had very low redundancy of information (i.e., each organization name usually containing only one word pointing to its category), and most of the information-bearing words occurred in the training sample only a few times. Moreover, the representation of text data spans much higher dimensional spaces than classical numerical datasets, for which the algorithms were initially developed. Statistical algorithms optimize locally, as there is no mathematical procedure for identifying global optima. In lower dimensional spaces, global optima can be determined through numerical approaches and repeated local optimization from different, randomly selected starting points (e.g. simulated

annealing). But in high dimensional spaces, due to limited computing power, it is not guaranteed that such methods will find true global optima (Gentzkow, Kelly, & Taddy, 2019).

For example, the rule-set developed by humans orders the search terms so that the „socialist student associations" are correctly classified, putting „socialist" high up in the hierarchy of tiers. The rule-set generated by machine learning, on the other hand, starts with those search terms that lead to the highest entropy reduction given the current viewpoint. If the search term „student union" reduces the entropy at one point by a higher amount than „socialist", then „student union" is put higher up in the hierarchy (or in other words: closer to the root node).

Discussion and recommendations for implementing a rule-based approach

An analysis of misclassified organizations reveals the limitations of the various approaches. The first major source of misclassifications is when NPOs carry names without surface information about their activities (e.g., an organization called „John Doe", or an uncommon acronym). Neither manual human coding, nor rule-based approaches, nor machine-learning approaches can classify such NPOs. In those cases, only acquiring additional information will help, e.g., humans doing desk research. However, the vast majority of NPOs carry names that contain at least some kind of information about their activities, even if they are not required to do so by law (as in the German case). The second major source of misclassifications, which poses problems for rule-based and machine learning approaches but not so much for manual human coding, is unconventional language use. This a problem with NPOs whose names include wordplay, neologisms, regional dialects and foreign languages (e.g., a choir called „coro.con.brio", a cats' shelter called *„Katzentant'"*).

Despite these obstacles, a rule-based approach using NPOs' names as input data delivers satisfactory results in circumstances where no longer high-quality texts about NPOs' are available. Rule-based approaches are semi-automated, i.e., experts with domain knowledge manually create a rule set for automatic classification. We have found that such a manually created rule-set is superior to rule-sets based on statistical algorithms if the quality of the input data is low. A comparison with a sample of German associations shows that these principles can also be applied to countries with relatively liberal laws on naming non-profit organizations. However, specific rule-sets are not transferable. They must be tailored to every civil society landscape. In the following sections, we give some concluding recommendations on how such rule-sets can be developed efficiently.

We implemented rule-based classification in the open-source software R. We provide the R script and the dictionary that form the rule-set under the conditions of a CC BY-NC-SA

4.0[6] license: [a link to the authors' university research depository will be provided here; for now, materials are provided through the editor to preserve the authors' anonymity]. The dictionary contains words, parts of words, and phrases that are related to an NPO's ICNPO category in one column. To integrate the dictionary into the rule-set, a second and third column are required. Those columns relate every single search term to an ICNPO category, and to a tier.

We recommend building the rule-set in the following way: The descriptions of the single ICNPO categories by Salamon and Anheier (1996) constitute a good starting point for the dictionary. All potential search terms (translated into the respective language) from those descriptions should be considered (e.g. „scouts"/ „*Pfadfinder*" as a term included in the description of the category „social services"). After testing for whether they deliver accurate results when applied to the full population of organizations, they can be included in the dictionary.

Subsequently, rules can be generated in a data-driven way. We wrote an additional short R script that calculates the frequency of all words in the names of all yet unclassified organizations. Based on this list, we worked our way down from the most frequent semantically significant words to less frequent ones. For each word we considered – and if it looked promising, we tried out – whether it was valid it on its own or as part of a search phrase to classify NPOs. Paying attention to linguistic details and carefully using truncation operators (such as .* as a placeholder for a flexible number of characters) turned out to be important in this process. For example, the search terms „*verband*" (association) or „*vertret.**" (represent/representing/represented/etc.) would not have worked as valid search terms on their own, but the phrase „*.*verband.*vertret.**" turned out to be a valid search term for identifying professional associations that should go into the category „business and professional associations, unions". Each potential search term needs to be tested by applying it to all yet unclassified associations. We also included search terms that deliver a high number of correct and a negligible number of incorrect classifications. Hence, we traded some classification error for higher overall classification rates.

When an organization's name is ambiguous, in the sense that it includes search terms pointing towards several ICNPO categories, those search terms need to be ordered hierarchically in

[6] This means that the materials may be used and adapted for non-commercial purposes, giving credit to us as authors and sharing adapted versions under the same conditions.

tiers. For example, our sample contained an organization called „sports union for people with disabilities", which includes terms pointing to the categories „sports" and „social services" (with recreation for people with disabilities explicitly mentioned as a case for the social services category in the guideline by Salamon and Anheier 1992). Because additional desk research showed this organization to be about competitive sports and hardly about providing social services, we assigned the search term „sports union" to a higher tier than „.*disabilities.*".

Within a tier, there are only search terms that do not appear together in an organization's name. So these tiers can also be understood as „OR" commands. Finding a suitable tier for a search term often involves some trial and error. A little programming detail helped to make this process more efficient: To facilitate cross-checking and correcting errors, we used preliminary ICNPO markers that include the tier on which the organization was classified, and generously added new tiers. If necessary, we added tiers in retrospect by re-assigning tier numbers with decimal places.

Performance can be improved by including wildcat term lists in the dictionary. These are lists of terms that are related to a particular concept. For example, the abstract concept of sport (for which there is an ICNPO category) manifests itself in many different kinds of sport. We used web scraping to obtain a list of over 200 officially recognized sports from an Austrian government website. We included those in a term list to assign organizations to the ICNPO category for sport. Similar approaches can be applied to generate lists of professions and jobs, medical and health-related terms, towns and regions, names of country citizens and ethnic groups, country names, and various kinds of animals. These wildcat term lists can be included in the dictionary like variables. For example, in the search term „friends.* country_adjective", the term „country_adjective" serves as a wildcat for the full list of countries in their adjective form (e.g., Armenian, Chinese …). With the use of such wildcat terms and search modifiers (especially the truncation operator .*), it is possible to build an elaborate and precise system of classification rules.

Researchers should be prepared that they will have to identify quite a large number of search terms, and that each search term by itself will only classify a relatively small number of cases (In our case, the strongest search term was „.**sparverein*.*", or in English „savings association", which led to the classification of 3.4% of the organizations in the sample.). The procedure for generating new search terms can be continued until the classification performance is satisfactory, or in theory indefinitely down to the level where new search

terms classify only one single NPO. The eventual rule-set in Austria contained a dictionary with 3090 search terms arranged in 211 tiers.

References

Fisher, I. E. G., M. R.;Hughes, M. E. . (2016). Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management, 23*(3), 157-214.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature, 57*(3), 535-574.

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.

Heitzmann, K., & Simsa, R. (2004). From corporatist security to civil society creativity: The nonprofit sector in Austria. In A. E. Zimmer & E. Priller (Eds.), *Future of Civil Society* (pp. 713-731): Springer.

Kwartler, T. (2017). *Text mining in practice with R*. Hoboken, NJ: Wiley.

Lantz, B. (2015). *Machine learning with R : discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R* (Second edition ed.). Birmingham Mumbai.

Lecy, J., & Thornton, J. (2016). What Big Data can tell us about government awards to the nonprofit sector. *Nonprofit and Voluntary Sector Quarterly, 45*(5), 1052-1069.

Lepere-Schloop, M. (2017). *The logic of identity-focused organizational change: Research based on the combined federal campaign.* (dissertation), University of Georgia, Retrieved from https://getd.libs.uga.edu/pdfs/lepere-schloop_megan_a_201708_phd.pdf

Lepere-Schloop, M., Zook, S., & Bawole, J. N. (2018). *NGO classification from the bottom-up: Using self-reported data and machine learning to generate categories of NGOs in Ghana*. Paper presented at the ISTR 13th International Conference, Amsterdam.

McDonnell, D., & Rutherford, A. C. (2018). The determinants of charity misconduct. *Nonprofit and Voluntary Sector Quarterly, 47*(1), 107-125.

Salamon, L. 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.

Salamon, L. M., & Anheier, H. K. (1996). *The International Classification of Nonprofit Organizations: ICNPO-Revision 1, 1996*. Retrieved from Baltimore Mar: Johns Hopkins University Institute for Policy Studies:

United Nations. (2018). Satellite Account on Non-profit and Related Institutions and Volunteer Work. Retrieved from https://unstats.un.org/unsd/nationalaccount/docs/UN_TSE_HB_FNL_web.pdf

Vandor, P., Traxler, N., Millner, R., & Meyer, M. (2017). *Civil society in Central and Eastern Europe: challenges and opportunities* (3902673109). Retrieved from http://epub.wu.ac.at/6256/1/Study_Civil-Society-in-CEE_WU-Wien.pdf

Zhai, C. X., & Massung, S. (2016). *Text data management and analysis: a practical introduction to information retrieval and text mining*. New York, NY: Association for Computing Machinery and Morgan & Claypool.

Zimmer, A., Gärtner, J., Priller, E., Rawert, P., Sachße, C., Strachwitz, R. G., & Walz, R. (2004). The legacy of subsidiarity: The nonprofit sector in Germany. In *Future of civil society* (pp. 681-711): Springer.