AUTOMATED CODING USING MACHINE-LEARNING AND REMAPPING THE U.S. NONPROFIT SECTOR: A GUIDE AND BENCHMARK

Ji MA

LBJ School of Public Affairs and RGK Center for Philanthropy and Community Service
The University of Texas at Austin

Nonprofit and Voluntary Sector Quarterly
https://doi.org/10.1177/0899764020968153

Abstract

This research developed a machine-learning classifier that reliably automates the coding process using the National Taxonomy of Exempt Entities as a schema and remapped the U.S. nonprofit sector. I achieved 90% overall accuracy for classifying the nonprofits into nine broad categories and 88% for classifying them into 25 major groups. The intercoder reliabilities between algorithms and human coders measured by kappa statistics are in the "almost perfect" range of 0.80–1.00. The results suggest that a state-of-the-art machine-learning algorithm can approximate human coders and substantially improve researchers' productivity. I also reassigned multiple category codes to over 439 thousand nonprofits and discovered a considerable amount of organizational activities that were previously ignored. The classifier is an essential methodological prerequisite for large-N and Big Data analyses, and the remapped U.S. nonprofit sector can serve as an important instrument for asking or reexamining fundamental questions of nonprofit studies. The working directory with all data sets, source codes, and historical versions are available on GitHub (https://github.com/ma-ji/npo_classifier).

Keywords: National Taxonomy of Exempt Entities, nonprofit organization, neural network, BERT, machine-learning, computational social science

---

Correspondence: 2315 Red River St, Austin, TX 78712, USA; maji@austin.utexas.edu. *Acknowledgment*: Earlier versions of this paper were presented at the 2019 West Coast Nonprofit Data Conference and 2019 ARNOVA Annual Conference. I thank the panel attendees and their constructive comments. I also thank Dr. Francie Ostrower, Dr. Brice McKeever, Dr. Pamela Paxton, Dr. Jesse Lecy, Dr. Mark Hager, Dr. Richard Steinberg, Dr. Diarmuid McDonnell, and Dr. Peter Frumkin for their constructive comments; Isha Kanani, Yizhuo Li, Meiying Xu, Li Ye, and Wenkang Li for their research assistance; Michael Shensky for the access to GIS resources. I thank the Texas Advanced Computing Center at the University of Texas at Austin for cloud computing resources. I thank the three anonymous reviewers, Dr. Michael Meyer, Dr. Angela Bies, and Dr. Viviana Wu for their constructive comments and handling this manuscript. *Funding information*: This study was supported in part by the 2019-20 PRI Award and Stephen H. Spurr Centennial Fellowship from the LBJ School of Public Affairs and a Planet Texas 2050 grant from UT Austin. *Biography*: Ji Ma is an Assistant Professor in Philanthropic and Nonprofit Studies at the Lyndon B. Johnson School of Public Affairs at the University of Texas at Austin. He studies and teaches state-society relationship, knowledge production, and computational social science methods from the perspective of nonprofit and philanthropy.

1 Introduction

Voluntary and philanthropic organizations have existed for centuries, but the term "nonprofit sector" was just coined in the 1970s by scholars and policy makers (Hall 2006). A major reason for assembling diverse organizations into a conceptual whole was to legitimize their existence and the benefits they received (54-55). As Barman (2013) pointed out, the order and structure of a society can be reflected by a classification system from Durkheim's perspective (Durkheim 2012). The National Taxonomy of Exempt Entities (NTEE) developed by the National Center for Charitable Statistics (NCCS) is the most widely used classification system and represents one of the efforts put forth to legitimize the existence of the nonprofit sector (Hodgkinson and Toppe 1991; Hodgkinson 1990). Since its creation, NTEE has been widely used in classifying nonprofits in the U.S. and as a benchmark for developing new classification systems. Methodologically, scholars also use NTEE as a coding schema to operationalize their primary constructs.

This research developed a machine-learning classifier that reliably automates the coding process using NTEE and remapped the U.S. nonprofit sector by reassigning multiple NTEE codes to organizations with purposes across various domains. The classifier is an essential methodological prerequisite for large-N and Big Data analyses, and the remapped U.S. nonprofit sector can serve as an important instrument for asking or reexamining fundamental questions of nonprofit studies. The working directory with all datasets, source codes, and historical versions are deposited on GitHub (https://github.com/ma-ji/npo_classifier). Although the progress made in this single study may not entirely solve all the challenges of NTEE, and this preliminary project can only serve as a stimulus for future studies, it provides an essential knowledge base and novel directions.

1.1 A short history of the NTEE classification system

In an effort to become legitimate, the development of the NTEE classification system dates from the 1980s (Hodgkinson 1990, 8-9, 11). In 1982, the NCCS assembled a team of experts who worked on creating a taxonomy for nonprofit organizations. The first draft of the NTEE schema came out in 1986 and was published in 1987. By the early 1990s, the NCCS had classified nearly one million nonprofits by using the NTEE. Then in 1995, the Internal Revenue Service (IRS) adopted the NTEE coding system, took over assigning and maintaining the classification, and started releasing the Business Master File with NTEE codes (US Internal Revenue Service 2014, 2013).

Two agencies were responsible for assigning these NTEE codes: the NCCS and the IRS. Before 1995, the NCCS coded nonprofits according to their program descriptions in Parts III and VIII of

Form 990, which were supplemented with information from Form 1023 ("Application for Recognition of Exemption") and additional research (National Center for Charitable Statistics 2006, 16). After 1995, the IRS began to issue "new exempt organizations an NTEE code as part of the determination process," and "the determination specialist [assigned] 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 [was] recognized as tax-exempt" (US Internal Revenue Service 2013, 1).

The NTEE classification system has supported many applied and academic studies on nonprofit organizations which have critical economic and political roles in society. For example, the NTEE provides a framework through which the social and economic activities of civil society can be mapped and compared with other social sectors (e.g., Roeger, Blackwood, and Pettijohn 2015). Scholars can use NTEE codes to sample nonprofits of interest (e.g., Okten and Weisbrod 2000; Sharkey, Torrats-Espinosa, and Takyar 2017; McVeigh 2006; Vasi et al. 2015) or as independent variables (Sloan 2009). The NTEE can also serve as an analytical tool to measure organizational capacity in different service domains and inform practitioners and policy makers (Hodgkinson and Toppe 1991). Moreover, scholars also use NTEE as a coding schema to operationalize their primary constructs (e.g., McVeigh 2006; Denison 2009; Bhati and McDonnell 2020)

1.2 Worst classification, except for all the others: Five problems of NTEE

The NTEE classification system, despite being one of the best we have so far, still has numerous critical drawbacks. First, because the NTEE only assigns one major category code to an organization, it cannot accurately describe a nonprofit's programs that are usually diverse and spread across several service domains (i.e., the *multi-code problem*; Grønbjerg 1994, 303). Even though a program classification system was later developed (Lampkin, Romeo, and Finnin 2001), it is still not widely used, probably because it is impractical to assign codes to a massive number of programs.

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" (i.e., the *incomplete information problem*; National Center for Charitable Statistics 2006, 16). A recent study found the number of organizations in Washington state with a specific NTEE code would increase significantly if mission statements were used for coding (Fyall, Moore, and Gugerty 2018).

Third, NTEE codes are static, whereas nonprofit organizations' activities may change over time (i.e., the *changing-code problem*). Recoding existing NTEE assignments is extremely onerous, and this may be one of the reasons that the IRS does not have a procedure through which non-

profits can request a change to their NTEE codes (US Internal Revenue Service 2013). The tremendous amount of human labor needed for classification is a prominent challenge and an obvious barrier to improving any classification system. This issue leads to the fourth *onerous labor problem*.

Fifth, a vast amount of grassroots organizations are not classified and remain missing in existing datasets because an organization "that normally has annual gross receipts of $50,000 or less" is not required to report to the IRS (i.e., the *missing-nonprofit problem*; US Internal Revenue Service 2019). As Smith (1997) estimates, the IRS listings ignore about 90% of nonprofits, most of which are grassroots associations. By surveying the communities in Indiana, Grønbjerg, Liu, and Pollak (2010, 931) found that about 40% of all the nonprofits in the state were not registered with the IRS. The nonprofits' activities at the grassroots level are particularly important, but many studies failed to consider these organizations because of the dataset limitation (e.g., McVeigh 2006; Vasi et al. 2015; Sharkey, Torrats-Espinosa, and Takyar 2017).

Numerous studies have experimented with computational methods in automating the coding process in research (e.g., Salminen et al. 2019; Baćak and Kennedy 2018; Nelson et al. 2018; Fyall, Moore, and Gugerty 2018; Anastasopoulos and Whitford 2019; Hollibaugh 2018), but many of these studies are introductory guides with showcases and are not solving a real-world research question. Bearing the five drawbacks in mind, I applied the advances in computational linguistics and contributed to the growing literature from these aspects: 1) I established a standardized workflow and benchmarks that future studies of nonprofits or typologies in other social science disciplines can build on and make comparisons to; 2) I achieved 90% overall accuracy for classifying the nonprofits into nine broad categories and 88% for classifying them into 25 major groups, and the intercoder reliabilities between algorithms and human coders measured by kappa statistics are in the "almost perfect" range of 0.80–1.00 (Landis and Koch 1977, 165); and 3) I solved the *multi-code problem* and remapped the U.S. nonprofit sector, which can serve as an important instrument for asking or reexamining fundamental questions of nonprofit studies. Ultimately, I developed a classifier that reliably automates the coding process using NTEE as a schema—an essential methodological prerequisite for large-N and Big Data analyses.

2 Method

Classifying nonprofits using their text descriptions is a typical task in automatic content analysis and usually employs three types of methods: the dictionary, supervised, and unsupervised methods (Grimmer and Stewart 2013, 268-269). The dictionary method uses a predefined dictionary of words to classify the texts. The automated classification method developed by the NCCS in

2007 belongs to this rule-based dictionary approach (The Nonprofit Center 2008). Two recent nonprofit classification studies also primarily adopted this approach (Fyall, Moore, and Gugerty 2018; Litofcenko, Karner, and Maier 2020). Although accurate and easy to implement, the dictionary approach cannot deal with the variations in and contexts of language. For example, "Hearts of Stone" is classified as "Housing Development, Construction, Management" because it matches the keyword "stone".[1] The supervised method is an improved solution that uses computer algorithms to learn the linguistic patterns in a dataset classified by human coders. Unlike the dictionary and supervised methods, which require predefined categories of interest, the unsupervised method can discover linguistic patterns in texts without inputting any knowledge of classification. However, the unsupervised method's validity can be problematic because the returned classifications may not be theoretically and practically meaningful. To take advantage of existing human-coded NTEE classifications and literature, this study employs a supervised approach as Figure 1 illustrates.

Figure 1 presents this paper's complete workflow. The ultimate goal of automated text classification is to devise a classifier that can replace robust human-coding. I implemented four stages of analysis to achieve this task: 1) the *preprocessing stage* included data acquisition and the preprocessing of datasets and texts[2]; 2) *feature extraction* included a bag-of-words representation (used by naïve Bayes and random forest algorithms) and word embedding (used by neural network algorithms); 3) the *training and decision-making* phase, was where I used stochastic and grid searches to train, search, and optimize the machine-learning algorithms; and 4) the last phase involved *training the model finalist* with the complete dataset and preparing the trained model for public use. Although the rest of this section introduces the four phases, this short article's focus is not to introduce detailed computational concepts and algorithms since they have been discussed in textbooks and aforementioned journal articles. Instead, I focus on how to apply these methods within nonprofit studies context.

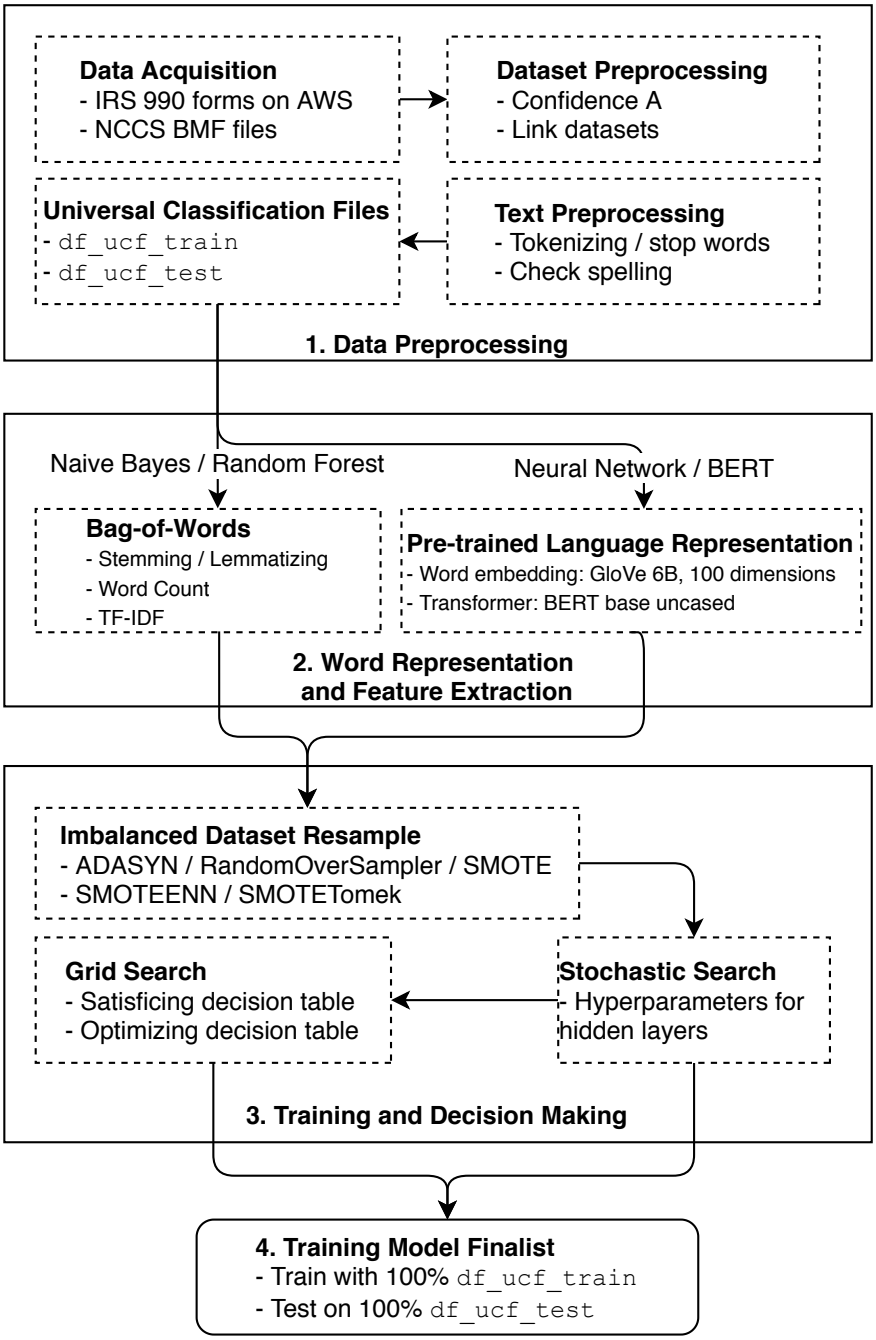
2.1 Data preprocessing

Data acquisition and dataset preprocessing. I collected text records from Forms 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 nonprofit organizations. Smaller 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, 1) can file Form

[1] Thank Dr. Mark Hager for this example.

[2] Although the classifier is developed using the texts from tax forms, it can also be used to classify other text documents.

Figure 1: RESEARCH WORKFLOW



990-EZ ("Short Form Return of Organization Exempt From Income Tax"), which is a shorter version of Form 990. Private foundations use Form 990-PF ("Return of Private Foundation"). The texts describe organizational activities in two forms: the overall mission statement and specific program descriptions. Table 1 summarizes these text fields' specific locations on the different forms.

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, Lines 2a-e, Lines 11a-c; Schedule O
990-EZ	Part III	Part III, Lines 28-30; Schedule O
990-PF	–	Part IX-A; Part XVI-B

Table 2: NTEE-CC CLASSIFICATION SYSTEM

Broad Category Code	Explanation	Major Group Code
I	Arts, Culture, and Humanities	A
II	Education	B
III	Environment and Animals	C, D
IV	Health	E, F, G, H
V	Human Services	I, J, K, L, M, N, O, P
VI	International, Foreign Affairs	Q
VII	Public, Societal Benefit	R, S, T, U, V, W
VIII	Religion Related	X
IX	Mutual/Membership Benefit	Y
X	Unknown, Unclassified	Z

Classification records (i.e., NTEE codes) were collected from the 2014–2016 Business Master Files on the NCCS website.[3] This study deals with two types of NTEE classifications: 10 broad categories and 26 major groups. Table 2 shows the relationship between the broad categories and major groups. A detailed list of the 26 major groups can be found through the IRS (2014).

The accuracy of a classification is indicated by the letters of A, B, and C, where a "confidence level of A ... indicates that there is at least a 90 percent probability that the major group classification is correct" (National Center for Charitable Statistics 2006, 16). From 2014 to 2016, 56.12% of records were classified at level A, 37.32% at level B, and 6.56% at level C. Records vary in confidence levels primarily because of information availability and clarity (The Nonprofit Center 2008). For example, a large amount of nonprofits have no mission statement and program description reported, so the NTEE codes for these organizations are assigned solely based on their names.

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

Only A-level records[4] are used for developing the algorithm. I made this decision because of three reasons. First, for training purposes, the training dataset needs to be of higher quality that includes typical features from which the algorithm can learn. Second, the intercoder reliability of records at confidence level A should approximate 100% (Stengel, Lampkin, and Stevenson 1998, 147). This measure is particularly important because the NTEE codes were assigned by human coders from different organizations (i.e., the IRS and NCCS) over different periods of time, and we need to assure that the reliability of the assigned codes are high. Third, about 1.76% of organizations changed their NTEE codes between 2014 and 2016. I excluded the records of these organizations because their category codes probably misrepresent their ongoing activities if they have not requested timely updates of their NTEE codes.

Selecting only A-level records does not undermine the reliability of the trained algorithm, but it has an important implication for future applications: When using our classifier, scholars should preprocess their text data to increase the quality before analysis. This is an essential step for any analysis.

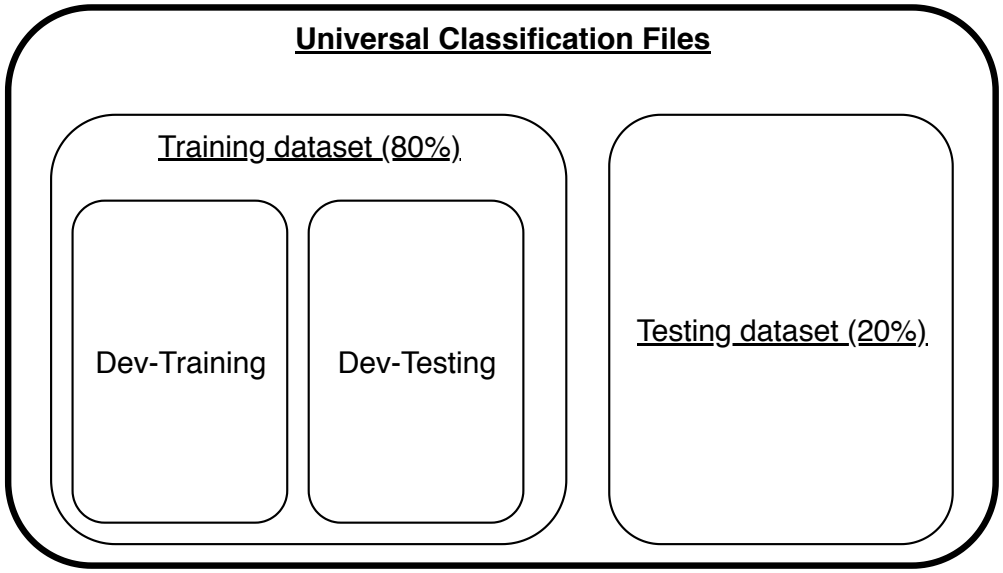
Text Preprocessing. Texts in sentences need to be segmented into words before analysis, which is called "tokenization" in natural language processing. For example, "we focus on education" needs to be tokenized into a list of words (i.e., "we," "focus," "on," "education"). I also removed stop words (e.g., "the," "a," "on," and punctuation marks) and checked spelling errors using algorithms based on "minimum edit distance" (i.e., the minimum number of editing operations needed to change one word into another; Jurafsky and Martin 2019, 23).

Universal Classification Files (UCFs). The final step in the data preprocessing stage is to divide data records into training and testing datasets (i.e., files in `/dataset/UCF/`) that are mutually exclusive and can be used to benchmark future models (Figure 2). The *Universal Classification File Training* (UCF-Training; `df_ucf_train.pkl.gz`) is used to develop models and comprises 80% of the total records. For developing models, the UCT-Training is also split into two mutually exclusive parts: training and testing subsets for developing algorithms. The *Universal Classification File Testing* (UCF-Testing; `df_ucf_test.pkl.gz`) is used to test a trained model's performance and comprises 20% of the total records. All records in UCF files are valid for training and testing purposes (i.e., all records have mission statement and program description information).

Table 3 presents the two datasets' composition by major groups. The UCFs approximate the composition of organizations reported to the IRS, except for groups *A* ("arts, culture, and humanities," more data) and *T* ("philanthropy, voluntarism, and grantmaking foundations," less data). The consequence is that the final algorithm's performance on *A* is more reliable because it was

[4] At this level, no records are in the *X/Z* category (i.e., unknown or unclassified).

Figure 2: STRUCTURE OF THE UNIVERSAL CLASSIFICATION FILES



trained with more data. However, the performance on category *T* is less reliable. So researchers using *T* organizations should be more cautious.

2.2 Word representation and feature extraction

After the data acquisition and preprocessing, we need to transform the tokenized sentences into numeric vectors used by the machine-learning algorithms. A variety of transformation methods can represent words as vectors, and good methods should be able to ease the process of extracting features from texts. In general, there are two approaches to word representation: bag-of-words and word embedding.

2.2.1 Bag-of-words approach

The bag-of-words approach considers words in texts as being mutually independent and thus disregards the order of the words. For example, "we are health service organization" and "health organization service are we" are the same from a bag-of-words perspective. This method serves as the basis for developing many simple language models because it can efficiently represent the possibility of a word's occurrence in texts (Bengfort, Bilbro, and Ojeda 2018). I adopted 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 that is

Table 3: COMPOSITION OF UNIVERSAL CLASSIFICATION FILES

Major Group	Training (#)	Training (%)	Testing (#)	Testing (%)	Reported (#)	Reported (%)
A	17,010	11.02%	4,291	11.11%	35,813	6.77%
B	25,827	16.72%	6,419	16.63%	67,879	12.83%
C	3,323	2.15%	827	2.14%	9,054	1.71%
D	4,239	2.75%	1,034	2.68%	8,740	1.65%
E	9,015	5.84%	2,307	5.98%	25,643	4.85%
F	2,301	1.49%	543	1.41%	8,481	1.60%
G	5,053	3.27%	1,353	3.50%	10,697	2.02%
H	467	0.30%	126	0.33%	2,203	0.42%
I	2,947	1.91%	740	1.92%	8,687	1.64%
J	4,772	3.09%	1,132	2.93%	15,841	2.99%
K	2,009	1.30%	522	1.35%	7,444	1.41%
L	5,942	3.85%	1,537	3.98%	20,428	3.86%
M	4,693	3.04%	1,140	2.95%	10,857	2.05%
N	15,460	10.01%	3,925	10.17%	43,987	8.31%
O	1,731	1.12%	409	1.06%	7,878	1.49%
P	9,180	5.94%	2,318	6.00%	40,880	7.73%
Q	1,987	1.29%	436	1.13%	7,288	1.38%
R	1,064	0.69%	257	0.67%	2,830	0.53%
S	14,459	9.36%	3,603	9.33%	48,387	9.14%
T	2,032	1.32%	541	1.40%	84,338	15.94%
U	1,000	0.65%	225	0.58%	3,039	0.57%
V	350	0.23%	85	0.22%	940	0.18%
W	8,357	5.41%	2,038	5.28%	20,862	3.94%
X	4,566	2.96%	1,098	2.84%	20,699	3.91%
Y	6,640	4.30%	1,701	4.41%	15,712	2.97%
Z	–	–	–	–	547	0.10%
Total	154,424	100.00%	38,607	100.00%	529,154	100.00%

Note: Numbers and percentages reported to the Internal Revenue Service (i.e., the last two columns) are from McKeever, Dietz, and Fyffe (2016). Dashed lines separate the 10 broad categories.

Table 4: EXAMPLE OF COUNT VECTORS

statement × vocabulary	we	focus	on	education	health	care	about
we focus on education	1	1	1	1	0	0	0
health care care	0	0	0	0	1	2	0
we care about	1	0	0	0	0	1	1

called the vocabulary index. The algorithm then represent the texts using word frequencies and the vocabulary index. Table 4 presents a simple example of count vectors in which the statement "we focus on education" is represented as the vector $[1,1,1,1,0,0,0]$.

Term frequency-inverse document frequency (TF-IDF) normalizes raw word frequencies using the number of documents in which a given 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 in which word i appears. The underlying assumption of TF-IDF is that any words appearing in all the statements are not as important as those occurring in a limited number of statements (Jurafsky and Martin 2019, 105).

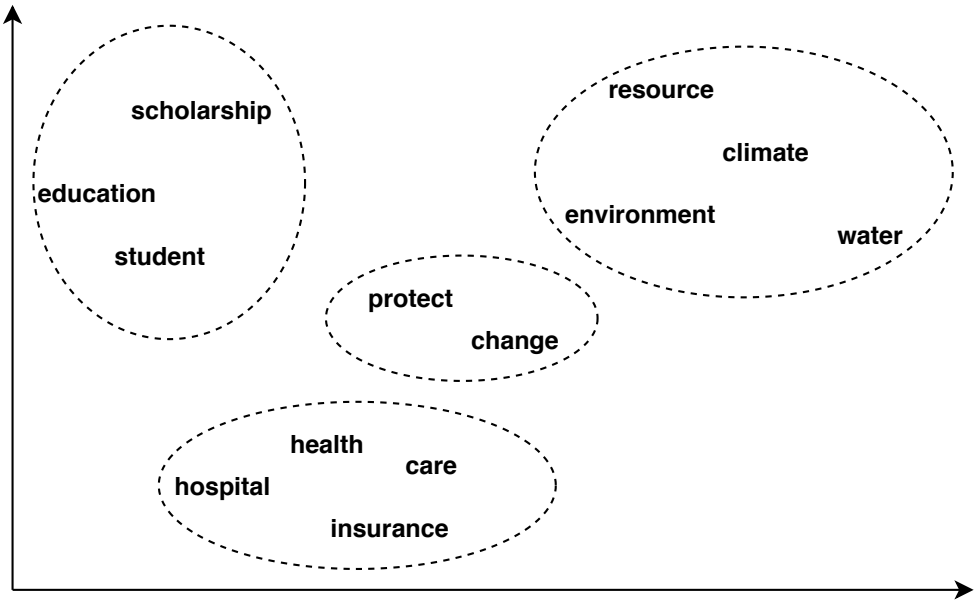
$$w_{ij} = tf_{ij} \cdot idf_i \tag{1}$$

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

We need to normalize the texts to reduce the vocabulary size before transforming them by 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 machine-learning models will suffer from "the curse of dimensionality": as the feature increases, the data become more discrete and less informative to decision making (Bellman 2015, 94).

The process of normalizing words is called "morphological parsing," which includes two primary methods: *stemming* and *lemmatizing* (Jurafsky and Martin 2019, 21). Stemming (i.e., "Porter Stemmer" in this study) slices longer strings into smaller ones according to a series of predefined rules. For example, "ational" is transformed to "ate" in all words ending with "ational." Therefore, stemming tends to have both over- and under-parsing errors. Lemmatizing (i.e., lemmatizer

Figure 3: WORD EMBEDDING EXAMPLES



based on WordNet in this study; Miller 1995) is a more advanced method that reduces a word to its stem with the help of part of speech tagging.

2.2.2 Pre-trained language representation approach

Disregarding the contexts in which the words appear is an evident flaw of the bag-of-words approach. Therefore, vectorizing words using a large text corpus (e.g., the entire English Wikipedia corpus) has become the basis for many state-of-the-art algorithms of natural language understanding. The word embedding approach is a new advancement (Mikolov et al. 2013) and was suggested by Nelson et al. (2018, 28) as a future direction for sociological studies, but it only has been applied by a few social scientists very recently (Kozlowski, Taddy, and Evans 2019). As Figure 3 illustrates, this method represents words in a multidimensional space (i.e., each word has a vector value), and words that often appear together in the text corpus are closer to each other (Jurafsky and Martin 2019, 99; Bengfort, Bilbro, and Ojeda 2018, 65). We can either train our own word vectors, which would require a large corpus and is time-consuming, or use pretrained word vectors. In this study, I used the 100-dimension word vectors pretrained from a corpus of 6 billion word tokens (Pennington, Socher, and Manning 2014). For the word embedding approach, we do not need to normalize the texts using stemming or lemmatizing because the dataset of pretrained word vectors contains all spelling variations, and the variations of the same word are close to each other in the multidimensional vector space.

Although this word embedding method can consider semantic contexts, it only gives fixed vector values to words; therefore, this method cannot handle the variations of context between tasks. The Bidirectional Encoder Representations from Transformers (BERT) is a newest solution to this issue (Devlin et al. 2019). The BERT model is first pre-trained using an unlabeled (i.e., unsupervised) corpus to have pre-trained parameters, and these parameters can be fine-tuned using the labeled corpus from downstream tasks. Simply put, the BERT model uses unlabeled large text corpus to obtain a range of values for different parameters, and then uses a specific task (i.e., classifying nonprofits in this paper) to fine-tune and get more accurate values.

2.3 Training and decision making

2.3.1 Imbalanced dataset resampling

Training using an imbalanced dataset such as UCF-Training can bias our prediction of minor classes because machine-learning algorithms cannot extract enough information from these classes (e.g., groups *H* and *V*). Therefore, resampling the imbalanced dataset to build a more balanced one is crucial for predicting minority classes. I experimented with three strategies of over-sampling (i.e., ADASYN, RandomOverSampler, and SMOTE) and two strategies of over-sampling followed by under-sampling to reduce the noise (i.e., SMOTEENN and SMOTETomek; Lemaître, Nogueira, and Aridas 2017). The influence of resampling is substantial: the F_1 score for predicting minority class major group *Q* was improved from 15% to over 30% in our pilot experiments.[5]

2.3.2 Classifiers for training

One principle of text analysis is that "there is no globally best method" (Grimmer and Stewart 2013, 270). For different tasks, it is important to test the performance of different families of classifiers. I experimented the typical models of four families: Naïve Bayes model based on probability theory, Random Forest model based on decision tree, convolutional neural network model based on deep neural network, and linear regression model. Because linear regression is familiar to most of the academic community, this section briefly introduces the first three models.

The *naïve Bayes (NB) classifier* is built on Bayes' theorem. It is one of the simplest classifiers to learn and implement among all machine-learning algorithms and is 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 and Martin 2019, 58; Grimmer and

[5] Although major group *Q* and broad category VI represent the same group of organizations, for computer algorithms, the classification contexts are different; therefore, performance on this category varies.

Stewart 2013, 277). I tested two types of NB classifiers: the multinomial and complement NB classifiers (Rennie et al. 2003).

The *random forest (RF) classifier* 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 make a prediction for the same record. A prediction class that is selected by most of these small algorithms is given as the prediction result by the RF algorithm. It uses the word "forest" because each small algorithm trained is a decision tree (Quinlan 1986, 83). A decision tree represents a set of questions that usually have yes/no answers. The process starts from the top of the tree (i.e., root node) with one question, and based on the answer, we run down either side of the tree and answer another question. We can repeat this process until reaching the end of the tree. Each decision tree is trained on a different training set (Breiman 1996, 124).

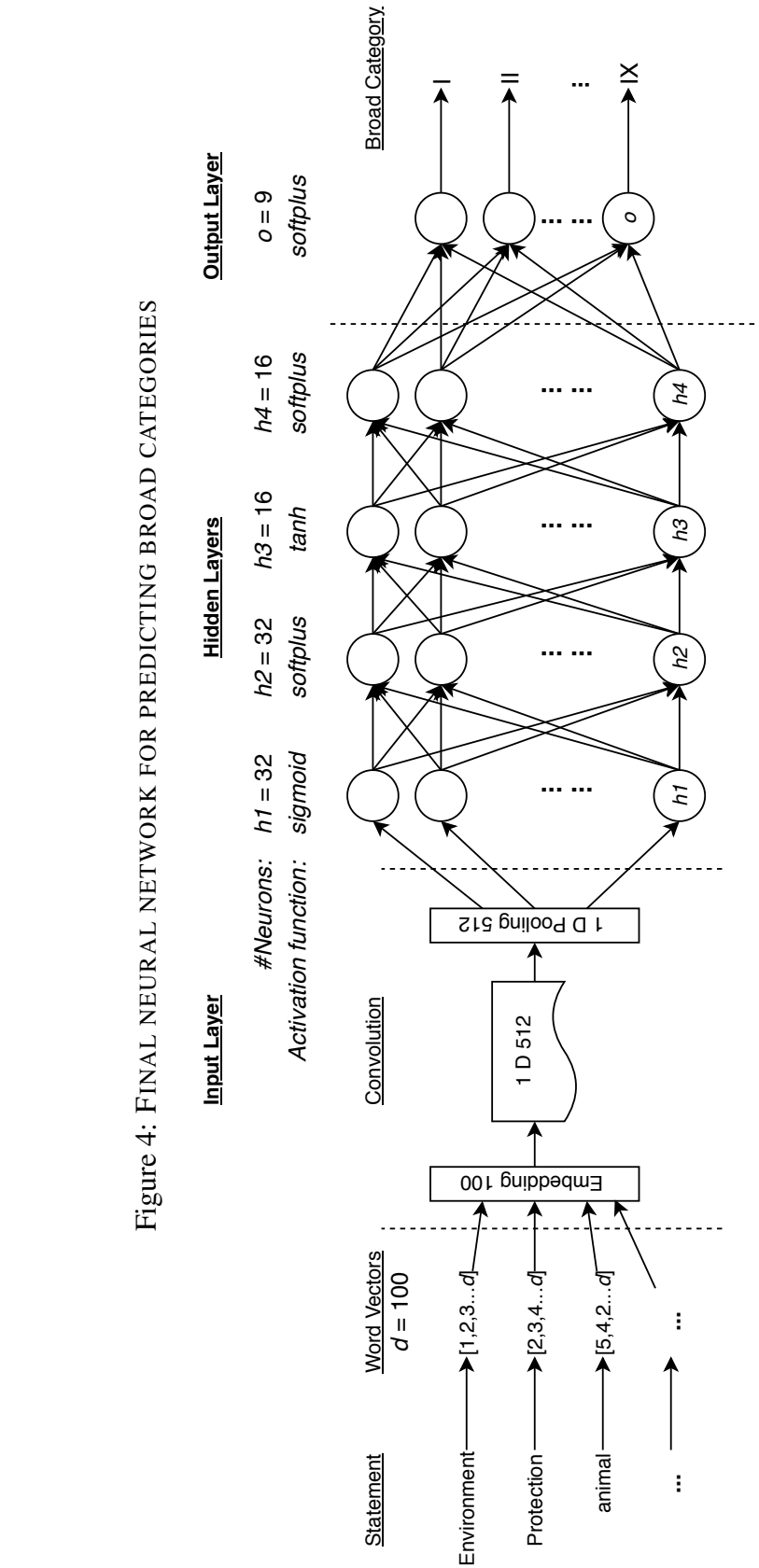
Neural network (NN) classification mimics the neural structures in human brains. Figure 4 illustrates the architecture of the final neural network for predicting broad categories. As the figure shows, each "neuron" (or node) is a simple classification function (e.g., a sigmoid or rectified linear unit function). We can arrange these neurons to form three types of layers (i.e., input, hidden, and output), and therefore, they can perform more complicated classification tasks. The connection between neurons has a numerical value called "weight." In the training stage, each neuron processes one record in a turn and learns by looking at the record's classification (i.e., the NTEE code) and comparing it with the known previous records. With every new record the neurons learn, they update the connection weight to update the model (Collobert and Weston 2008, 163). After the network is done processing each record in 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, or long short-term memory). This study uses convolutional NN (CNN) following scholars' recommendation (Zhang and Wallace 2015).

The classifiers use different approaches to vectorize words: the NB and RF classifiers use the bag-of-word approach, the NN classifier employs the word embedding approach, and the BERT classifier is a pre-trained BERT embedding (i.e., not fine-tuned bare BERT embedding) with a layer of linear regression nodes on top.

2.3.3 Measuring algorithm performance

An algorithm's performance can be measured by many metrics, but social scientists are particularly concerned with three questions when solving real-word problems: 1) How many predicted observations are correct (i.e., *precision* calculated by Eq. 3)? 2) How many observations are cor-

Figure 4: FINAL NEURAL NETWORK FOR PREDICTING BROAD CATEGORIES



rectly predicted (i.e., *recall* calculated by Eq. 4)? 3) How reliably can algorithms replace human coders (i.e., *intercoder reliability*)? Answering these three questions is critical for social scientists who apply machine-learning research methods. Moreover, instead of using terms that are only familiar to computer scientists, I introduce these measures with the NTEE classification contexts.

In Eq. 3, k is one of the NTEE codes, $\#Org_k^{pred}$ is the number of organizations predicted as k (i.e., the sum of *true positive* and *false positive*), and $\#Org_k^{corr}$ is the number of correct predictions (i.e., *true positive*). $\#Org_k^{corr}$ will always be smaller than or equal to $\#Org_k^{pred}$ because machine-learning algorithms can hardly predict every observation correctly. For example, $Precision_B = 0.75$ indicates that 75% of all the organizations classified as "education" are correct.

$$Precision_k = \frac{Org_k^{corr}}{Org_k^{pred}} \tag{3}$$

In Eq. 4, Org_k^{human} is the number of organizations that belong to k category robustly coded by a human (i.e., the sum of *true positive* and *false negative*). For example, $Recall_B = 0.80$ denotes that 80% of the organizations classified as "education" by robust human coding are correctly identified by the algorithm.

$$Recall_k = \frac{Org_k^{corr}}{Org_k^{human}} \tag{4}$$

The precision and recall are competitive; that is, the increase of one measure will sacrifice the other. Therefore, the F_1 score (Eq. 5), the harmonic mean of precision and recall, was introduced to balance the two measures.

$$F_{1k} = \frac{2 \cdot Precision_k \cdot Recall_k}{Precision_k + Recall_k} \tag{5}$$

We can also calculate the *intercoder reliability* between an ML algorithm and a human coder since our ultimate goal is to use the former to replace the latter. The kappa-type statistics are a widely used measure of intercoder reliability, and Landis and Koch (1977, 165) provided the following interpretation: less than 0, poor; 0.00–0.20, slight; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; 0.81–1.00, almost perfect. Simundic et al. (2009) used these statistics to compare human coders and automated methods in biomedicine and achieved scores at the "moderate" range. However, the interpretations of the kappa measure is suggestive. Whether the value is sufficient also depends on the research question (Viera and Garrett 2005).

2.3.4 Decision making

The goal of this study is to find the best machine-learning algorithm from an extensive collection of parameters. We can either try some of the configurations randomly (i.e., *stochastic search*) or iterate all possible configurations (i.e., *grid search*). For the NB and RF algorithms, I used the latter approach. For the NN algorithms, I first used a stochastic search to narrow the configurations of hidden layers and then conducted a grid search for the input and output layers' parameters using a CNN. The grid search for all possible parameters (over 2 million combinations) is impossible even when using one of the most advanced supercomputing clusters in the world.

I conducted two rounds of grid searches. The first round was for *satisficing decision making* in which I only considered the configurations that can perform at the top 5% (240 parameter combinations for NB and RF and 7,200 for NN, detailed history files are in the `output/` folder). Then I ran the second found grid search for *optimizing decision making* in which I increased the values of some parameters to allow the algorithms to reach their performance ceilings. I then chose the best algorithm and parameters for final training.

3 Results

3.1 Selecting the model with best performance

For the multiclass classification task (i.e., more than two classes to predict), it is difficult to measure the overall performance because the performance differs for each category. Table 5 presents the performance of the CNN classifiers with and without resampling. Because the dataset is imbalanced, the classifier performs poorly on category VI ("international, foreign affairs") without resampling. Training the classifier with a resampled dataset substantially improved the F_1 score from 14% to 29% but slightly sacrifices its performance on other categories. So which one should we choose?

I chose the classifier trained without resampling as the best model because even though the F_1 score for category VI was substantially improved, we could not use the predicted results for this category (21% identified of which only 44% are correct). I recommend not sacrificing the performance on other categories since researchers need to manually check or completely drop this category in their analysis anyway. For social scientists, mathematical improvements may not yield substantial and practical meanings. This rationale applies to selecting other classifiers.

Table 5: COMPARING CONVOLUTIONAL NEURAL NETWORK CLASSIFIERS

Code	Precision-*N*	Precision-*R*	Recall-*N*	Recall-*R*	F_1-*N*	F_1-*R*	%Obs.
I	87%	83%	85%	87%	86%	85%	11%
II	85%	91%	88%	78%	86%	84%	17%
III	76%	83%	90%	82%	82%	82%	5%
IV	76%	88%	87%	70%	81%	78%	11%
V	85%	77%	86%	90%	85%	83%	30%
VI	59%	44%	8%	21%	14%	29%	1%
VII	88%	83%	76%	79%	81%	81%	17%
VIII	65%	71%	77%	70%	71%	71%	3%
IX	90%	80%	85%	92%	88%	85%	4%

Note: *N* = No resampling; *R* = Resampling.

3.2 Performance of the best model

After experimenting four classifiers with extensive parameters, the fine-tuned BERT classifier has the best performance: For classifying the nine broad categories, 90% of records in the UCF-Testing dataset were correctly recognized, and the intercoder reliability kappa measure is 0.88; for the 25 major-group task, 88% were correctly classified, and their kappa measure is 0.87. Both kappa statistics are in the "almost perfect" range (i.e., between 0.80 and 1.00; Landis and Koch 1977, 165). The values of precision, recall, and F_1 for each category and group varies, as presented in Tables 6 and 7.[6]

Our BERT classier outperformed human coders on many broad categories (i.e., I, III, V, VII, and IX; five out of nine) and major groups (i.e., *A*, *D*, *G*, *H*, *J*, *K*, *M*, *N*, *R*, *S*, *T*, *V*, *W*, and *Y*; 14 out of 25). For example, the classifier outperformed human coders on broad category VII ("public, societal benefit"): 88% of the category VII organizations were identified, and among these identified organizations, 90% were correct—14% higher than the human coders' performance. For major group *W* ("public, society benefit - multipurpose and other"), 94% of the group *W* organizations were identified, and among these identified organizations, 92% were correct—34% higher than human coders. A caveat is, Stengel, Lampkin, and Stevenson (1998) did the verification 20 years ago. The data quality at that time was probably inferior to what it is now. However, the algorithm in this study is trained and tested using high-quality records (i.e., A-level records). Therefore, it may not be surprising that the algorithm outperforms human coders.

All the predicted results are generally satisfactory, except for the major group *V* ("social science research institutes"). This group had the poorest performance: only 48% of the group *V* organi-

[6]Tables A1 and A2 in the appendix have more measures of performance. These measures are not widely employed by the machine learning community but presented here to compare with results from elsewhere.

Table 6: PERFORMANCE OF BEST MODEL ON BROAD CATEGORY

NTEE	HP	NB	CNN	BERT
I	88	82-86-84	87-85-86	92-92-92
II	93	84-82-83	85-88-86	91-91-91
III	87	77-86-81	76-90-82	90-92-91
IV	92	76-81-78	76-87-81	90-88-89
V	86	83-81-82	85-86-85	90-92-91
VI	77	25-76-38	59-8-14	67-68-68
VII	76	83-73-78	88-76-81	90-88-89
VIII	87	73-55-63	65-77-71	82-84-83
IX	90	86-83-84	90-85-88	91-94-92

Notes: Numbers show percentages (Precision-Recall-F_1). NTEE = National Taxonomy of Exempt Entities; HP = Human coder precision, compiled from Stengel, Lampkin, and Stevenson (1998, 153); NB = Naïve Bayes; CNN = Convolutional Neural Network; BERT = Bidirectional Encoder Representations from Transformers.

zations were identified, and among these identified organizations, only 59% were correct. Even though the precision is 35% higher than a human coder's precision, the predicted values cannot directly be used in analysis. Researchers should be cautious if their research questions are related to "social science research institutes." The low human and algorithmic precision may also suggest that the construct validity of this major group is questionable, which can be a direction for future studies.

3.3 Remapping the U.S. nonprofit sector

I solved the *multi-code problem* and remapped the U.S. nonprofit sector using the trained classifier. For each organization, the classifier returns a raw score for each NTEE code. The raw scores (i.e., in machine learning terms, "logits") are between $(-\infty, +\infty)$, but we can normalize them to probabilities (i.e., values between $[0, 1]$) using either a softmax or sigmoid function. The softmax function treats all categories as mutually exclusive, and the sum probability of all NTEE codes is equal to 1. While the sigmoid function treats all classifications as independent, and the sum probability is not constrained to 1. Therefore, the sigmoid transformation can help us solve the *multi-code problem*.

I validated the predicted results by manually checking a sample of 200 observations (i.e., confidence interval $95\% \pm 7\%$). Among the 200 records, 10.5% of them have incomplete informa-

Table 7: PERFORMANCE OF BEST MODEL ON MAJOR GROUP

NTEE	HP	NB	CNN	BERT
A	88	87-82-84	80-87-83	93-92-92
B	93	85-78-81	85-85-85	92-91-91
C	86	62-77-69	65-74-69	82-86-84
D	90	87-89-88	80-90-85	92-94-93
E	92	77-69-73	77-78-78	87-85-86
F	86	59-55-57	51-60-55	77-77-77
G	65	65-56-60	68-68-68	83-86-84
H	73	33-56-41	55-19-28	81-63-71
I	84	63-64-63	71-71-71	84-85-85
J	72	71-77-74	86-67-75	86-81-84
K	82	68-67-67	63-68-66	84-84-84
L	83	68-71-70	70-76-73	83-84-83
M	88	81-84-82	87-90-88	93-94-93
N	88	87-86-87	83-93-88	94-95-94
O	91	58-52-55	65-61-63	83-84-84
P	88	56-62-59	64-57-60	75-78-76
Q	77	47-53-50	43-36-39	67-67-67
R	67	39-56-46	46-21-28	74-69-72
S	75	75-77-76	84-79-81	90-88-89
T	78	43-47-45	66-32-43	83-67-74
U	76	27-46-34	52-22-31	67-78-72
V	24	0-0-0	0-0-0	59-48-53
W	58	87-80-84	87-86-86	92-94-93
X	87	63-74-68	68-71-70	81-85-83
Y	90	82-88-85	84-91-88	91-94-92
Z	10		–	

Notes: Numbers show percentages (Precision-Recall-F_1). NTEE = National Taxonomy of Exempt Entities; HP = Human coder precision, compiled from Stengel, Lampkin, and Stevenson (1998, 153); NB = Naïve Bayes; CNN = Convolutional Neural Network; BERT = Bidirectional Encoder Representations from Transformers. Dashed lines separate the ten broad categories.

tion.[7] For the remaining records, the accuracy of IRS-reported NTEE is 87.71%. For the predicted NTEE codes, I split the results into three groups: high (normalized probability $\geq$ .99), medium ($\geq$ .95), and low ($\geq$ .90). Manual validation revealed that the accuracy for these categories is 91.94%, 83.33%, and 62.82%, respectively. The cumulative accuracy is 89.39% (i.e., high + medium) and 83.33% (i.e., high + medium + low). I decided to combine the results from the high and medium categories because it can provide more classification labels and outperform the accuracy of IRS-reported NTEE.

Figure 5 illustrates the remapped U.S. nonprofit sector in comparison to the original classifications registered with IRS, and an online interactive visualization can be accessed at https://jima.me/?ntee_remap. Disparities are many, but the most substantial change is the reduced percentage of *T* ("philanthropy, voluntarism, and grantmaking foundations"). The NCCS assigned *T* to all private foundations without examining their purposes. This wild approach assumed that these foundations "[make] grants to unrelated organizations or institutions or to individuals" (National Center for Charitable Statistics 2007, 13). This coding criterion can be useful to avoid "double-counting" (Hodgkinson 1990, 17), but many of these private foundations clearly specified their service areas and could be operational foundations (i.e., not distributing grants to other nonprofits). Therefore, the current *T* category registered with IRS is significantly inflated and cannot reflect the actual activities because it is assigned by institutional type but not organizational purposes. The remapped U.S. nonprofit sector can provide a more accurate description and serve as an important instrument for asking or reexamining fundamental questions of nonprofit studies.

3.4 Python package for classifying texts

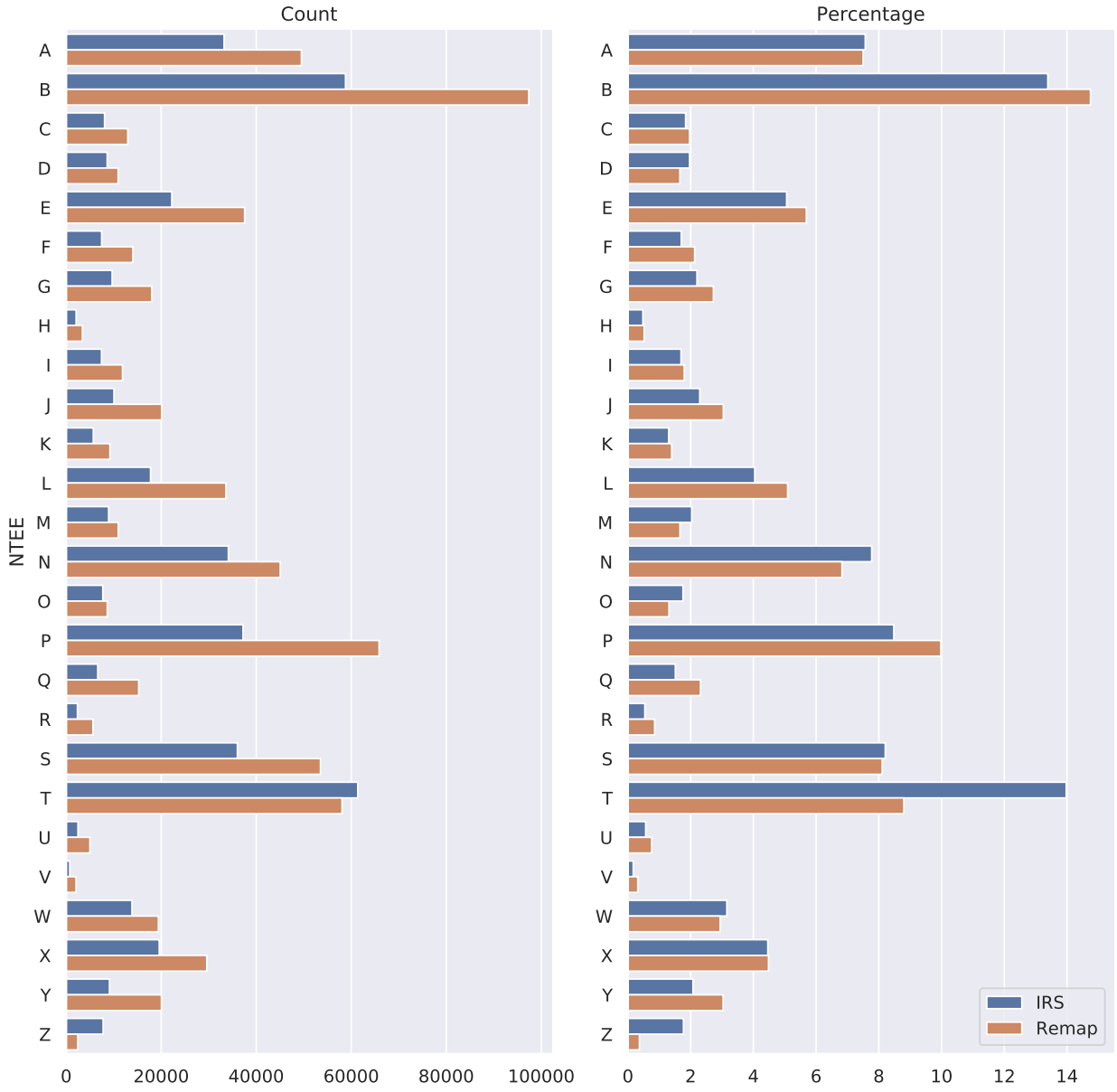
I developed a Python package (`npoclass`, under folder `npo_classifier/API`) for classifying texts using NTEE codes, and scholars can use it free of charge. Although the package was developed using the texts from tax forms, researchers can also use it to classify other text documents, for example, program and fundraising descriptions, news articles, and automated scraped websites. But like all analytical tasks, the raw text data need to be carefully preprocessed, as I have introduced in the method section. The package's documentation has more instructions.

4 Discussion

I developed a classifier that can reliably automate the coding process using NTEE as a schema—an essential methodological prerequisite for large-N and Big Data analyses. I achieved 90% over-

[7] A record is treated as "incomplete" if its organization name, mission statement, and program description cannot provide meaningful information for inferring the NTEE code.

Figure 5: REMAPPING THE U.S. NONPROFIT SECTOR



Note: NTEE = National Taxonomy of Exempt Entities; IRS = Internal Revenue Service. Using 990 Forms of the 439,160 nonprofit organizations that reported to the IRS in 2018 (https://registry.opendata.aws/irs990/).

all accuracy for classifying the nonprofits into nine broad categories according to their text descriptions, and 88% for classifying them into 25 major groups (both excluding the category of "unknown"). The intercoder reliabilities between algorithms and human coders (i.e., NTEE values coded by humans in NCCS Business Master Files) measured by kappa statistics are in the "almost perfect" range (i.e., between 0.80 and 1.00; Landis and Koch 1977, 165). I solved the *multi-code problem* and remapped the U.S. nonprofit sector by reassigning multiple NTEE codes to organizations with purposes across various domains. In general, an encouraging takeaway of this study is that machine-learning algorithms can approximate human coders and substantially improve a researcher's productivity, and the remapped U.S. nonprofit sector can serve as an important instrument for asking or reexamining fundamental questions of nonprofit studies.

4.1 On the way to conquering the five problems

This paper may not conquer all the problems of NTEE introduced earlier, but it provides an essential knowledge base and novel directions for future studies. The classifier alone is not a sufficient solution, but it is a powerful tool to make all the problems solvable.

Although the primary challenge, the *multi-code problem*, is solved directly, the remapping may cause double-counting if scholars use multiple NTEE codes in their research because one organization is counted in different categories. However, the double-counting issue is not specific to the devised classifier, but to all multi-label classification systems. Depending on research question, this issue may bias estimation.

For the *incomplete information problem*, first, I used more available information in our classification (Table 1) than existing studies that only used titles, mission statements, and program descriptions in Part III of the 990 forms. Second, for the organizations that only have limited information on the 990 forms or do not file tax forms at all (e.g., unincorporated grassroots voluntary groups), we can generate information from elsewhere. For example, one of our ongoing projects has retrieved the names, descriptions, and comments of thousands of grassroots organizations and groups through Google Map API. We then classified these information using the devised classier and had a more holistic picture of the nonprofit sector in a certain metropolitan area. This solution also applies to the *missing-nonprofit problem*.

An important takeaway from this project is that, even though information scarcity is not the most severe issue now, the ability to process information is much more challenging. The *changing-code* and *onerous labor problems* also result from a lack of information processing ability. This paper enables us to tackle these challenges with confidence. Although preliminary, it established a benchmark for future work.

4.2 Practical suggestions to social scientists solving real-world problems

The performance results in this paper indicate that social scientists who want to apply computational methods in their research should be cautiously confident. The key notion supporting our confidence here is a robust validation because too many factors can influence the validity of the algorithm (Grimmer and Stewart 2013, 271). For example, the algorithm may perform poorly on a dataset that is structurally different from the training dataset. I strongly suggest that readers review the annotations in the scripts posted online to understand the caveats and then make necessary optimizations according to their own research questions.

Social scientists should also take advantage of high-performance computing (HPC) research infrastructures (e.g., Keahey et al. 2018). These machine-learning algorithms can achieve their best performance only when trained with a large amount of data, and such a training process consumes a huge amount of computing power that is far beyond the capacity of the most advanced personal computers. At the grid search phase of this study, I used two of the most advanced GPU accelerators (NVIDIA Tesla P100) for NN training and six 48-CPU computing servers for NB and RF training. The HPC infrastructures are widely used in natural sciences but are still new to social scientists. Methodology workshops should incorporate the introduction of HPC infrastructures into their syllabi.

Applications of this study are broad. For example, computational social scientists can apply the workflow presented in this paper to other domains of inquiry. Other than academic purposes, practitioners can also use our study for industrial purposes. For example, classifying program descriptions and matching volunteering interests. Future studies can make numerous improvements based on the workflow and benchmark introduced in this paper. First, studies on this topic can experiment with more classifiers and parameters, for example, applying a more accurate nonprofit-specific glossary and stemmer (Paxton, Velasco, and Ressler 2019). Second, I deposited the working directory with all datasets, source codes, and historical versions on GitHub, enabling future large-scale collaborations on this project. A competition event on this subject is also being prepared.[8] Third, the Python software package can be improved with the inputs from scholars. Last but not least, we are advancing a multilingual version of this project using the International Classification of Nonprofit Organizations (Salamon and Anheier 1992) to assist the study of nonprofits in non-English-speaking countries. This step is essential for studying global civil society (Vakil 1997; Salamon and Anheir 1996).

[8] https://jima.me/?npo-classifier-competition

REFERENCES

Anastasopoulos, L. Jason, and Andrew B. Whitford. 2019. "Machine Learning for Public Administration Research, With Application to Organizational Reputation." *Journal of Public Administration Research and Theory* 29, no. 3 (June 7, 2019): 491–510.

Baćak, Valerio, and Edward H. Kennedy. 2018. "Principled Machine Learning Using the Super Learner: An Application to Predicting Prison Violence." *Sociological Methods & Research* (January 10, 2018): 0049124117747301.

Barman, Emily. 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, Richard E. 2015. *Adaptive Control Processes, A Guided Tour.* Princeton: Princeton University Press.

Bengfort, Benjamin, Rebecca Bilbro, and Tony Ojeda. 2018. *Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning.* 1 edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly Media, July 1, 2018.

Bhati, Abhishek, and Diarmuid McDonnell. 2020. "Success in an Online Giving Day: The Role of Social Media in Fundraising." *Nonprofit and Voluntary Sector Quarterly* 49, no. 1 (February 1, 2020): 74–92.

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24, no. 2 (August 1, 1996): 123–140.

Collobert, Ronan, and Jason Weston. 2008. "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning." In *Proceedings of the 25th International Conference on Machine Learning,* 160–167. ICML '08. New York, NY, USA: ACM.

Denison, Dwight V. 2009. "Which Nonprofit Organizations Borrow?" *Public Budgeting & Finance* 29, no. 3 (September 1, 2009): 110–123.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding" (May 24, 2019).

Durkheim, Émile. 2012. *The Elementary Forms of the Religious Life.* Courier Corporation.

Fyall, Rachel, M. Kathleen Moore, and Mary Kay Gugerty. 2018. "Beyond NTEE Codes: Opportunities to Understand Nonprofit Activity Through Mission Statement Content Coding." *Nonprofit and Voluntary Sector Quarterly* 47, no. 4 (August): 677–701.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–297.

Grønbjerg, Kirsten 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, no. 3 (October 1, 1994): 301–328.

Grønbjerg, Kirsten A., Helen K. Liu, and Thomas H. Pollak. 2010. "Incorporated but Not IRS-Registered: Exploring the (Dark) Grey Fringes of the Nonprofit Universe." *Nonprofit and Voluntary Sector Quarterly* 39, no. 5 (October 1, 2010): 925–945.

Hall, Peter Dobkin. 2006. "A Historical Overview of Philanthropy, Voluntary Associations, and Nonprofit Organizations in the United States, 1600–2000." In *The Nonprofit Sector: A Research Handbook,* edited by Walter W Powell and Richard Steinberg, 32–65. Yale University Press.

Hodgkinson, Virginia A. 1990. "Mapping the Non-Profit Sector in the United States: Implications for Research." *Voluntas: International Journal of Voluntary and Nonprofit Organizations* 1, no. 2 (November 1, 1990): 6–32.

Hodgkinson, Virginia A., and Christopher Toppe. 1991. "A New Research and Planning Tool for Managers: The National Taxonomy of Exempt Entities." *Nonprofit Management and Leadership* 1 (4): 403–414.

Hollibaugh, Gary E. 2018. "The Use of Text as Data Methods in Public Administration: A Review and an Application to Agency Priorities." *Journal of Public Administration Research and Theory.*

Jurafsky, Daniel, and James H. Martin. 2019. *Speech and Language Processing.* 3rd draft. October 16, 2019.

Keahey, Kate, Pierre Riteau, Dan Stanzione, Tim Cockerill, Joe Mambretti, Paul Rad, and Paul Ruth. 2018. "Chameleon: A Scalable Production Testbed for Computer Science Research." In *Contemporary High Performance Computing: From Petascale toward Exascale,* 1st ed., edited by Jeffrey Vetter, vol. 3. Chapman & Hall/CRC Computational Science. Boca Raton, FL: CRC Press.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings." *American Sociological Review* 84, no. 5 (October 1, 2019): 905–949.

Lampkin, Linda, Sheryl Romeo, and Emily Finnin. 2001. "Introducing the Nonprofit Program Classification System: The Taxonomy We've Been Waiting for." *Nonprofit and Voluntary Sector Quarterly* 30, no. 4 (December 1, 2001): 781–793.

Landis, J. Richard, and Gary G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33 (1): 159–174.

Lemaître, Guillaume, Fernando Nogueira, and Christos K. Aridas. 2017. "Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *J. Mach. Learn. Res.* 18, no. 1 (January): 559–563.

Litofcenko, Julia, Dominik Karner, and Florentine Maier. 2020. "Methods for Classifying Nonprofit Organizations According to Their Field of Activity: A Report on Semi-Automated Methods Based on Text." *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations* 31, no. 1 (February 1, 2020): 227–237.

McKeever, Brice S., Nathan E. Dietz, and Saunji D. Fyffe. 2016. *The Nonprofit Almanac: The Essential Facts and Figures for Managers, Researchers, and Volunteers.* Rowman & Littlefield, October 12, 2016.

McVeigh, Rory. 2006. "Structural Influences on Activism and Crime: Identifying the Social Structure of Discontent." *American Journal of Sociology* 112, no. 2 (September 1, 2006): 510–566.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space" (January 16, 2013).

Miller, George A. 1995. "WordNet: A Lexical Database for English." *Commun. ACM* (New York, NY, USA) 38, no. 11 (November): 39–41.

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

National Center for Charitable Statistics. 2007. *National Taxonomy of Exempt Entities-Core Codes 2007 Desk Reference.* Washington, DC: The Urban Institute.

Nelson, Laura K., Derek Burk, Marcel Knudsen, and Leslie McCall. 2018. "The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods." *Sociological Methods & Research* (May 27, 2018): 0049124118769114.

Okten, Cagla, and Burton A. Weisbrod. 2000. "Determinants of Donations in Private Nonprofit Markets." *Journal of Public Economics* 75, no. 2 (February): 255–272.

Paxton, Pamela, Kristopher Velasco, and Robert Ressler. 2019. "Nonprofit-Specific Glossary and Stemmer." Accessed May 9, 2019. https://web.archive.org/web/20190509160945/https://www.pamelapaxton.com/990missionstatements.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "Glove: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 1532–1543. Doha, Qatar: Association for Computational Linguistics, October.

Quinlan, J. R. 1986. "Induction of Decision Trees." *Machine Learning* 1, no. 1 (March 1, 1986): 81–106.

Rennie, Jason D. M., Lawrence Shih, Jaime Teevan, and David R. Karger. 2003. "Tackling the Poor Assumptions of Naive Bayes Text Classifiers." In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning,* 616–623. ICML'03. AAAI Press.

Roeger, Katie L., Amy S. Blackwood, and Sarah L. Pettijohn. 2015. "The Nonprofit Sector and Its Place in the National Economy." In *The Nature of the Nonprofit Sector,* Third edition, edited by J. Steven Ott and Lisa A. Dicke, 22–37. Boulder, CO: Westview Press, July 28, 2015.

Salamon, Lester M, and Helmut K. Anheier. 1992. "In Search of the Non-Profit Sector II: The Problem of Classification." *Voluntas: International Journal of Voluntary and Nonprofit Organizations* 3, no. 3 (December 1, 1992): 267–309.

Salamon, Lester M, and Helmuth K Anheir. 1996. *The International Classification of Nonprofit Organizations ICNPO-Revision 1, 1996.* Baltimore, Md: The Johns Hopkins University Institute for Policy Studies.

Salminen, Joni, Vignesh Yoganathan, Juan Corporan, Bernard J. Jansen, and Soon-Gyo Jung. 2019. "Machine Learning Approach to Auto-Tagging Online Content for Content Marketing Efficiency: A Comparative Analysis between Methods and Content Type." *Journal of Business Research* 101 (August 1, 2019): 203–217.

Sharkey, Patrick, Gerard Torrats-Espinosa, and Delaram Takyar. 2017. "Community and the Crime Decline: The Causal Effect of Local Nonprofits on Violent Crime." *American Sociological Review* 82, no. 6 (December 1, 2017): 1214–1240.

Simundic, Ana-Maria, Nora Nikolac, Valentina Ivankovic, Dragica Ferenec-Ruzic, Bojana Magdic, Marina Kvaternik, and Elizabeta Topic. 2009. "Comparison of Visual vs. Automated Detection of Lipemic, Icteric and Hemolyzed Specimens: Can We Rely on a Human Eye?" *Clinical Chemistry and Laboratory Medicine* 47 (11): 1361–1365.

Sloan, Margaret F. 2009. "The Effects of Nonprofit Accountability Ratings on Donor Behavior." *Nonprofit and Voluntary Sector Quarterly* 38, no. 2 (April 1, 2009): 220–236.

Smith, David Horton. 1997. "The Rest of the Nonprofit Sector: Grassroots Associations as the Dark Matter Ignored in Prevailing "Flat Earth" Maps of the Sector." *Nonprofit and Voluntary Sector Quarterly* 26, no. 2 (June 1, 1997): 114–131.

Stengel, Nicholas A. J., Linda M. Lampkin, and David R. Stevenson. 1998. "Getting It Right: Verifying the Classification of Public Charities in the 1994 Statistics of Income Study Sample." In *Turning Administrative Systems Into Information Systems,* edited by Statistics of Income Division and Internal Revenue Service, 6:145–167. Statistics of Income Division, Internal Revenue Service.

The Nonprofit Center. 2008. "How and by Whom Are NTEEs Assigned?," May 28, 2008. Accessed January 18, 2020. https://web.archive.org/web/20200118035322/http://www.thenonprofitlink.org/knowledgebase/detail.php?linkID=728&category=120&xrefID=3012/.

US Internal Revenue Service. 2013. "IRS Static Files No. 2013-0005," March 29, 2013. Accessed November 27, 2018. https://www.irs.gov/pub/irs-wd/13-0005.pdf.

US Internal Revenue Service. 2014. "Exempt Organizations Business Master File Information Sheet," April. Accessed November 27, 2018. https://www.irs.gov/pub/irs-soi/eo_info.pdf.

US Internal Revenue Service. 2018. "2017 Instructions for Form 990-EZ," January 29, 2018. https://www.irs.gov/pub/irs-pdf/i990ez.pdf.

US Internal Revenue Service. 2019. "Annual Exempt Organization Return: Who Must File." Accessed May 8, 2019. https://www.irs.gov/charities-non-profits/annual-exempt-organization-return-who-must-file.

Vakil, Anna C. 1997. "Confronting the Classification Problem: Toward a Taxonomy of NGOs." *World Development* 25, no. 12 (December 1, 1997): 2057–2070.

Vasi, Ion Bogdan, Edward T. Walker, John S. Johnson, and Hui Fen Tan. 2015. ""No Fracking Way!" Documentary Film, Discursive Opportunity, and Local Opposition against Hydraulic Fracturing in the United States, 2010 to 2013." *American Sociological Review* 80, no. 5 (October 1, 2015): 934–959.

Viera, Anthony J., and Joanne M. Garrett. 2005. “Understanding Interobserver Agreement: The Kappa Statistic.” *Family Medicine* 37, no. 5 (May): 360–363.

Zhang, Ye, and Byron Wallace. 2015. “A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification” (October 13, 2015).

A Appendix: Tables

Table A1: MORE PERFORMANCE MEASURES OF THE BEST MODEL: BROAD CATEGORIES

	Specificity	Geometric Mean	Index Balanced Accuracy
I	99.03%	95.30%	90.15%
II	98.31%	94.50%	88.63%
III	99.47%	95.41%	90.30%
IV	98.74%	93.47%	86.46%
V	95.72%	93.72%	87.49%
VI	99.62%	82.52%	65.96%
VII	98.03%	93.00%	85.64%
VIII	99.45%	91.14%	81.73%
IX	99.57%	96.50%	92.57%
Average	97.76%	93.87%	87.49%

Table A2: MORE PERFORMANCE MEASURES OF THE BEST MODEL: MAJOR GROUPS

	Specificity	Geometric Mean	Index Balanced Accuracy
A	99.08%	95.42%	90.39%
B	98.39%	94.60%	88.83%
C	99.59%	92.53%	84.46%
D	99.78%	96.80%	93.15%
E	99.19%	91.66%	82.79%
F	99.67%	87.70%	75.18%
G	99.37%	92.30%	84.04%
H	99.95%	79.16%	60.33%
I	99.69%	92.12%	83.63%
J	99.62%	89.73%	79.01%
K	99.78%	91.81%	83.01%
L	99.31%	91.08%	81.66%
M	99.77%	96.99%	93.56%
N	99.31%	97.07%	93.81%
O	99.82%	91.36%	82.12%
P	98.34%	87.36%	74.74%
Q	99.63%	81.96%	65.02%
R	99.84%	83.16%	67.03%
S	98.94%	93.43%	86.36%
T	99.80%	81.94%	64.96%
U	99.78%	88.09%	75.90%
V	99.92%	69.43%	45.71%
W	99.54%	96.51%	92.59%
X	99.42%	91.67%	82.78%
Y	99.56%	96.77%	93.13%
Average	99.15%	93.30%	86.26%