Classifying tax-exempt organizations with BERT and donation network information

Sophia Skowronski
sski@pm.me

Abstract

This report explores the use of network information to improve the automated labeling of National Taxonomy of Exempt Entities codes of tax-exempt organizations in the United States. Combining the processed data from Ma (2021) with 1.4 million grants parsed from 990 forms, we achieve 89.8% of weighted accuracy on 9 broad category groups 86.8% of the 25 major groups using a fine-tuned, pre-trained BERT model. More experimental results indicate that incorporating network data into the text directly, which summarizes donor behavior and location, gives limited improvements in the NLP classification task. The working directory with data download instructions and source codes is available on GitHub (`https://github.com/sophiaski/graph_network_ntee_classification`).

Keywords: National Taxonomy of Exempt Entities, nonprofit organization, BERT, social networks, text classification

1 Introduction

The National Taxonomy of Exempt Entities (NTEE) code is used by the Internal Revenue Service (IRS) to classify tax-exempt organizations. A specialist from the IRS assigns this as part of the review process when the organization is recognized as tax-exempt.

In practice, these labels represent nonprofit activity across a variety of domains. and due to the fact that these activities evolve over time and their taxonomies may no longer be relevant, the data may be noisy. It appears, though that these challenges are not insurmountable, as seen in prior work (Ma, 2021). Transformer architectures can approximate human coders in classifying these codes, and, moreover, this classifier can be leveraged to help automate the labeled of previously unassigned organizations and detect other philanthropic activity in new context.

2 Methods

Because the NTEE codes map to a broader, summary-level category label (Table 1), there are two classification tasks to optimize.

2.1 Data

The data for this work is sourced from a pre-processed data set from the benchmark paper on this same subject (Ma, 2021). It contains NTEE codes from the Exempt Organizations Business Master File Extract (BMF) and texts from IRS 990 Forms between 2014 and 2016. To extract the text for modeling, we concatenate the taxpayer name (i.e., organization name), mission statement texts, and all program description texts. The author sampled from a larger set of data based on human labeling accuracy scores received from the National Center for Charitable Statistics, only including organizations with the highest confidence levels. We maintain the same training and test split for better comparisons.

For both of the experiments defined later on, this data is combined with another data set of nonprofit grants made between 2010 and 2016. This enables us to see, given an employer identification number (EIN), what types of grants were received and given in this time frame by/to the organization and its region.

In summary, both data sources originate from IRS 990 forms, but their content, time frame, and parsing scheme vary.

Table 1: The NTEE score is used by the Internal Revenue Service and NCCS to classify tax-exempt organizations into 26 major groups that map onto 10 broad categories. We've excluded the "Unknown, Unclassified" categories from the modeling.

Broad category	Name	Major group
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
IV	Mutual/Membership Benefit	Y
X	Unknown, Unclassified	Z

2.2 Experimental Set-Up

Class Imbalance

As this is a multi-class classification problem, it is not surprising that there is class imbalance present in the data. Ignoring this would lead to a heavy bias towards the larger classes, where classes with fewer data points are treated as noise and are ignored. Therefore, accuracy metric is not as relevant when evaluating the performance of a model trained on imbalanced data.

To manage class imbalance when training the model, we will take two approaches. First, when we split between the training and validation data, we will ensure the same distribution of class labels is present in either set to make hyperparameter tuning more representative of the validation (and test) labels. Additionally, during the training phase, we will apply the weighted sampling in the data-loader modules to over-sample minority classes and under-sample majority classes with replacement. We do not want the training batches to contain samples from only a few of the classes that are over-represented. Ideally, each training batch will contain a uniform spread of the labels.

The evaluation metrics we will use are weighted f1-score and weighted accuracy to take into consideration the number of instances per label. Because we are primarily interested in labels that are correct by class and overall, these evaluation metrics work best for this set-up.

Baseline: Fine-tuned BERT

As a baseline for comparing experiments, we spun up the Bert for Sequence Classification pipeline to fine-tune the model for multi-class classification, using the existing pre-trained model from the Hugging Face Transformer library, as well as Adam optimization and Cross Entropy loss, both of which are common for multi-class problems.

Experiments

To confirm the assumptions made in the original paper, namely, that the organization data not included was prone to missing fields or mislabeled NTEE codes, two changes to the underlying data were made.

The first experiment sought to answer whether adding more data from previous years could improve the model, given that the difference in years between the two sets of data (2014-16 versus 2010-16) meant there was potentially three years worth of organization labels and text that could be integrated into the benchmark. Specifically, we identified new organizations not included in the benchmark, found their most up-to-date NTEE codes, cleaned their text fields, and added their data into the training and validation splits. Even if the programmatic and mission statements were not present to provide better quality text, it was worth exploring whether adding new text and label pairs could at least make improvements by requiring less re-balancing of the data.

The second experiment was a different type of data transformation: enhancing the text fields with time-based network information this are relevant to nonprofit activity: grant descriptions, organization type, and locations. In IRS 990 forms, charitable entities must disclose their top donations from a given year with descriptions on its purpose. As a result, it is likely that these descriptions provide more insight into the types of programs that non-profits are operating, which could be learned to provide better accuracy on the predicting the NTEE class. In order to include this information, we grouped the grants data by EIN and transferred the list of grants into a

Table 2: Optimal BERT hyperparameters

Parameter	Broad category	Major group
Learning rate	5×10^{-5}	3×10^{-5}
Classifier dropout	0.3	0.4
% warm-up steps	0.2	0
Max length	128	128
Epochs	3	3
Training batch size	32	32

more human-readable format and appended it to the end of the organization sequences. Similarly, organization type (grantor or grantee) with the set of operating locations was appended to the text string.

3 Results

See Table 3 for a summary of all experimental results.

3.1 Baseline Results

To find the most performant BERT model, a randomized sweep against an array of high dimensional hyperparameter spaces was initiated for both the broad and major category groups. Optimal parameters are listed in Table 2. This was an efficient way to make direct comparison of models and optimize by weighted accuracy, and use these features as a starting point for additional experimentation. As the weighted F1-scores did not vary too widely from weighted accuracy, we continued to use weighted accuracy as the optimization metric.

Of note, for the broad category label, we linearly increase the learning rate from a low rate to a constant rate after 20% of training steps are taken. This reduces volatility in the early stages of training. Beyond that, all of the parameters chosen were within common limits for the BERT Sequence Classification transformer architecture.

The benchmark paper claims 90% accuracy on the broad category and 88% on the major group, however it is inadvisable to use overall accuracy when the labels are imbalanced. I included their reported data points of the "Index Balanced Accuracy" with asterisks. Their F1 scores by label were calculated similarly so they are included in the table as well.

3.2 Experiment Results

In order to make the initial comparisons with baseline, we used the same BERT hyperparameters on each transformed data set. Note that each experiment would likely need their own hyperparameter sweep and fine-tuning to be comprehensive, but time constraints limited further model comparisons.

Adding new organizations and labels into the data, without further processing and IRS 990 form text extraction, leads to lower scores, even if it initially helped with class imbalance during training. By using only organization names as input, it is likely that by themselves are not good summary representations of nonprofit activity, especially given the limitations of the NTEE code itself. Longer text field descriptions of mission and programs, as reported in individual 990 forms, is necessary to attend to the correct context for predictions.

Secondly, adding network information such as donation history, location information, and organization roles as human-readable strings into the organization's text field provides a modest bump in scores for the broad classification model. The motivation for appending this data to the original sequence was to provide additional context for generalizing, especially for sequences that are shorter in length or have labels that are more universal. Case in point: for the major group classification task, adding network complexity increased the weighted accuracy and f1 scores for the minority labels. You can see further illustration of the calibration of the classifier in the confusion matrix (subsection 3.2).

Figure 1: Confusion matrix for broad category label. "International, Foreign Affairs" (6th column) has the smallest amount of true positives across the data set, with "Mutual/Membership Benefit" (7th column) having the largest. This validates the scores provided in the experimental results table. WeightsBiases interactive chart: `https://tinyurl.com/confusion-matrix-ntee-broad`

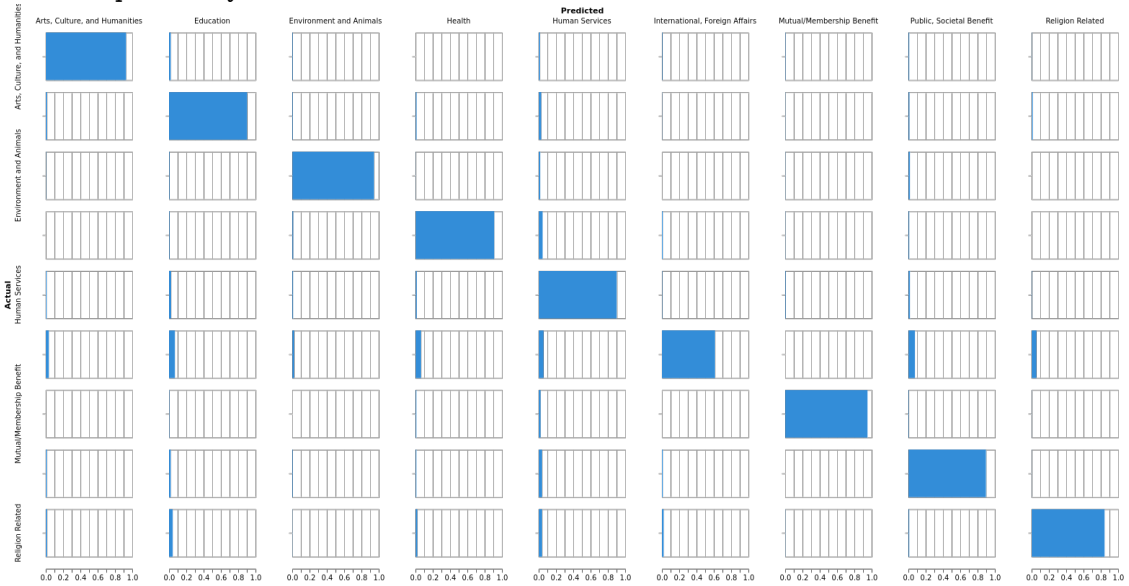

Table 3: Experimental results

Experiment$_{Broad}$	Acc$_{total}$	Acc$_{best(Mem)}$	Acc$_{worst(Int)}$	F1$_{total}$	F1$_{best(Mem)}$	F1$_{worst(Int)}$
BERT	**89.56**	94.9	58.25	**89.54**	91.01	61.75
BERT$_{+Targets}$	79.79	92.11	53.43	79.81	87.07	50.49
BERT$_{+Network}$	**89.85**	93.52	56.42	**89.81**	91.53	61.04
BERT$_{Ma,2021}$	**87.49***	92.57*	65.96*		89	68
Experiment$_{Major}$	Acc$_{total}$	Acc$_{best(Ani)}$	Acc$_{worst(SocSci)}$	F1$_{total}$	F1$_{best(Ani)}$	F1$_{worst(SocSci)}$
BERT	86.84	94.79	**51.43**	86.97	93.62	**52.17**
BERT$_{+Network}$	86.84	93.63	**58.57**	86.96	90.9	**56.94**
BERT$_{Ma,2021}$	86.26*	93.15*	**45.71***		93	**53**

4 Conclusion

This report was a useful exercise is validating the assumptions from nonprofit researchers, such as Fyall (2018), that mission and program statement text offers a better information source for predicting nonprofit activity than NTEE scores. Even more so, we may be able to make headway on this classification problem if the data focuses on what an organization does rather than the type of organization it is. The improvement on scores on minority classes when adding contextual information to the text supports this, however, more experimentation is necessary. Regardless, a fully-mapped United States nonprofit sector can serve as an important benchmark for evaluating fundamental questions and trends in the sector.

IRS 990 forms are a rich source of information for identifying how individuals (contractors, leadership, board members), organizations (grantors, grantees), nonprofit activity (grants), and regions are connected in philanthropic networks. As such, this work lays the f oundation for further exploration of modeling philanthropic activity as social networks, which will continue in the referenced GitHub repository.

References

Causebot.(2017). Nonprofit Grants 2010 to 2016 [Data set]. data.world. https://data.world/causebot/grant-2010-to-2016

Chen, H., Zhang, R. Identifying Nonprofits by Scaling Mission and Activity with Word Embedding. Voluntas (2021). https://doi.org/10.1007/s11266-021-00399-7

Fyall, R., Moore, M. K., Gugerty, M. K. (2018). Beyond NTEE Codes: Opportunities to Understand Nonprofit Activity Through Mission Statement Content Coding. Nonprofit and Voluntary Sector Quarterly, 47(4), 677–701. https://doi.org/10.1177/0899764018768019

Ma, J. (2021). Automated Coding Using Machine Learning and Remapping the U.S. Nonprofit Sector: A Guide and Benchmark. Nonprofit and Voluntary Sector Quarterly, 50(3), 662–687. https://doi.org/10.1177/0899764020968153