Detecting Informal Data References in Academic Literature

Sara Lafia[1], Jeong-Woo Ko[3], Elizabeth Moss[1], Jinseok Kim[2,3], Andrea Thomer[3], and Libby Hemphill[1,3]

[1]ICPSR, University of Michigan
[2]ISR, University of Michigan
[3]School of Information, University of Michigan

Abstract

The Inter-university Consortium for Political and Social Research (ICPSR) is developing a machine learning approach using natural language processing (NLP) to assist in the detection of informal data references. Formal data citations that reference unique identifiers are readily discoverable; however, informal references indicating research data reuse are challenging to infer and detect. We contribute a model that uses a combination of cues, such as the presence of indicator terms and syntactical patterns, to assign a likelihood score to dataset mentions and extract candidate data citations from academic text. In production, the model will support the evaluation of candidate documents for ingest into the ICPSR Bibliography of Data-related Literature. This work supports a larger effort to measure the impact of research data.

Keywords: data citation, data reference, machine learning, research data metrics

1 Introduction

Assessing the impact of research data requires knowledge of who has used data and for what purposes. Despite investments made to support research data preservation and curation, relatively little is known about how data are reused. Recent work investigating the relationship between properties of datasets, curation decisions, and reuse has found that curated data are used more often [3] but there is more to learn about the context surrounding data reuse.

The Inter-university Consortium for Political and Social Research (ICPSR) is a large social sciences data archive, which curates research data and maintains a collection of publications determined to have utilized data available at ICPSR. The ICPSR Bibliography of Data-related Literature[1] has strict criteria for inclusion and preserves only references to resources that indicate actual data reuse rather than mentions of datasets [10]. ICPSR staff manually curate the Bibliography, which is time-consuming given the volume of candidate citations returned for thousands of research studies in the archive.

2 Background

Initiatives to measure data impact, such as Project COUNTER [2], rely on formal data citation using persistent, unique identifiers (PIDs). The use of PIDs to reference datasets is an emerging practice, however [9, 12]. A study of the Dryad digital repository found that the share of articles referencing PIDs had grown from 69% to 83% between 2011 and 2014; however, the share of articles that included data identifiers in the works cited section remained low, under 10% [8]. Many researchers reference data informally, for example by study name in sections of the main text such as methods, as well as in footnotes, tables, acknowledgements, and supplements [12].

Until a culture of formal data citation is established, studies of research data impact will rely on detection of informal data citations. Machine learning (ML) and natural

[1] https://www.icpsr.umich.edu/web/pages/ICPSR/citations/

language processing (NLP) have been used to support bibliometric research, for example, to match citation strings to complex research objects, like longitudinal studies [7], and infer research fields and methods from citations [6].

We use ML and NLP to detect informal references to research data in the full text of academic publications. To test our approach, we use data from the Coleridge Initiative, a multi-year effort to demonstrate how publicly-funded data are used to inform decisions [5]. The data were made available as part of the Show US the Data[2] Kaggle competition, which was active from May to July 2021. We describe the task of detecting informal data references and its application at ICPSR, as well as our work's implications for the development of impact metrics for research data.

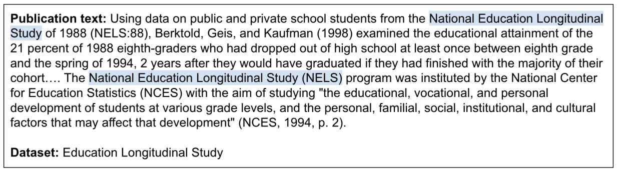
3 Approach

We used supervised and semi-supervised learning to predict whether a piece of text contained a reference to a dataset. This approach relied on input features to make a prediction. To train our models, we used a combination of heuristics defined by the ICPSR Bibliography team as input features [11] including whether the text contained an acronym, an indicator phrase, and was in a particular section of an article. We used the features to train a high recall random forest (RF) classifier to predict sentences that were likely to contain dataset references. We then trained a named entity recognition (NER) model to detect and extract informal dataset labels from candidate sentences. A combination of RF and NER models supported detection of informal references to research data in literature.

3.1 Training corpus

The training corpus was provided by the Coleridge Initiative competition and contained 14,271 unique, full text articles with references to 45 distinct datasets. The training data included publication identifiers, canonical data titles, and data labels that indicated the portion of the text where the data was referenced. In most instances, data labels were alternative titles used to refer to datasets; for example, data from the "Alzheimer's Disease Neuroimaging Initiative" was often referred to as the "ADNI database".

For the purposes of the competition, all training labels were considered true data references, meaning that they indicated actual data reuse. We tokenized the corpus text into 6,734,263 sentences using the Python Natural Language Toolkit (NLTK) [1]. Fewer than 1% of all sentences contained a data reference. Figure 1 illustrates a snippet of full publication text and labeled references to datasets.



Publication text: Using data on public and private school students from the National Education Longitudinal Study of 1988 (NELS:88), Berktold, Geis, and Kaufman (1998) examined the educational attainment of the 21 percent of 1988 eighth-graders who had dropped out of high school at least once between eighth grade and the spring of 1994, 2 years after they would have graduated if they had finished with the majority of their cohort.... The National Education Longitudinal Study (NELS) program was instituted by the National Center for Education Statistics (NCES) with the aim of studying "the educational, vocational, and personal development of students at various grade levels, and the personal, familial, social, institutional, and cultural factors that may affect that development" (NCES, 1994, p. 2).

Dataset: Education Longitudinal Study

Figure 1: Publication text with highlighted data references

3.2 Feature selection

We explored the features of sentences containing data labels for indicators of data references. First, we searched for a modified set of terms and phrases that were previously found to indicate data sharing and reuse [12]. Frequent terms included ".com", ".edu", "obtained from", "database", "survey", and others shown in Figure 2; the terms "deposit", "accession", and "donate" were not present. Next, we inspected the sections of articles where data references occurred. Prior studies had found that data references were commonly made in the methods or acknowledgements sections of research articles [12]. However, many data references included in the training data occurred in the introduction, abstract, or discussion sections of articles. Finally, we checked if the data reference sentences contained an acronym or a known Data.gov or ICPSR dataset title. We found that 33% of data references contained an acronym, while 1.9% contained a federal government dataset title listed on Data.gov and 0.26% contained an ICPSR dataset title. This method relied on string matching, so variations of titles were not captured.

3.3 Citation prediction

We trained a classifier using the Python scikit-learn library [13] to predict target sentences containing data labels. We formatted the features for our classifier as follows: number

[2] https://www.kaggle.com/c/coleridgeinitiative-show-us-the-data

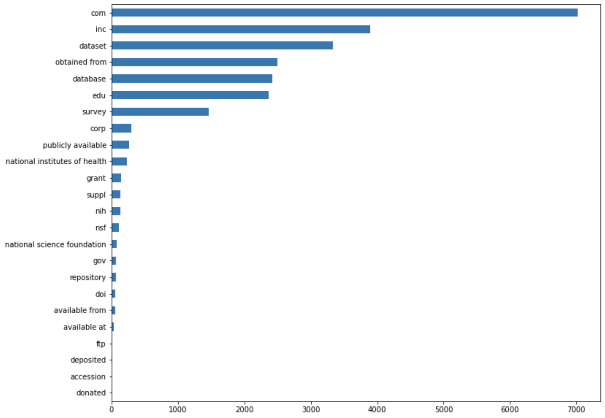


Figure 2: Counts of indicator terms in the training data

of indicator terms, number of acronyms, section information, and matching Data.gov or ICPSR titles. We trained and compared a dummy classifier, a logistic regression model, and random forest model using cross validation. Given the imbalanced class ratio of the training data (only 1/120 sentences contained data references), we set class weight to be inversely proportional to the class frequency. The random forest (RF) model with balanced class weights performed comparably to the logistic regression model with respect to accuracy (Table 1) but had a higher recall score of 0.99. We selected the balanced RF model to capture all candidate sentences containing data references.

Table 1: Classifier performance metrics

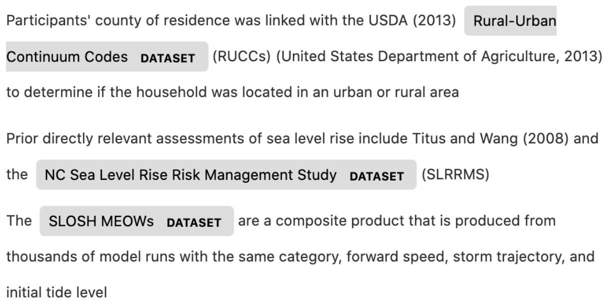
	Accuracy	**Balanced accuracy**
Dummy classifier	0.984	0.501
Logistic regression (LR)	0.993	0.678
Random forest (RF)	0.993	0.678
LR with balanced class weights	0.989	0.988
RF with balanced class weights	0.990	0.970

The following features of the balanced RF model were most predictive: the data label included a Data.gov title; had an acronym; and mentioned the indicator terms "data", "national" or "survey". We applied the classifier to the 2,722 test data sentences and generated prediction probabilities for each sentence. We discarded 99.4% of sentences that had a target probability score below 0.9, which reduced the total number of sentences to evaluate.

3.4 Entity extraction

We trained a natural language processing (NLP) function to detect a custom "dataset" entity using the Python spaCy library, which uses a default transformer architecture configuration [4]. For training, we tokenized a random subset of 500 articles into sentences. We sampled sentences for positive and negative examples of data references; the positive examples included data labels while negative examples did not. We downsampled the negative examples to address the imbalance between classes. We preprocessed the text to remove punctuation, leaving letters and digits. We then matched each training data label with each sentence and encoded the text span of the starting and ending characters for each matching entity in a "DocBin" format.

Figure 3 shows an example of output sentences with dataset entities highlighted. A notebook demonstrating our full approach is publicly available [3].



Participants' county of residence was linked with the USDA (2013) Rural-Urban Continuum Codes DATASET (RUCCs) (United States Department of Agriculture, 2013) to determine if the household was located in an urban or rural area

Prior directly relevant assessments of sea level rise include Titus and Wang (2008) and the NC Sea Level Rise Risk Management Study DATASET (SLRRMS)

The SLOSH MEOWs DATASET are a composite product that is produced from thousands of model runs with the same category, forward speed, storm trajectory, and initial tide level

Figure 3: Sentences with highlighted dataset entities

4 Evaluation

Competition submissions were evaluated using a Jaccard similarity measure comparing prediction strings to a ground truth. Our team's baseline solution[4] received our highest public F-score (0.49). This solution searched for

[3] https://www.kaggle.com/saralafia/v3-rf-ner

[4] https://www.kaggle.com/saralafia/v2-string-matching

training dataset labels and titles matching the test data and added an extended set of federal government dataset titles obtained from the CKAN API for Data.gov[5]. Our solution detailed in Section 3 received a lower public F-score (0.425); it used a classifier to predict which sentences contained data references and then used an NER model to extract candidate dataset entities. Both of our solutions had lower private F-scores, suggesting sizable differences in the composition of the public and private test corpora.

5 Conclusion

We plan to expand the ICPSR Bibliography through the detection of informal data references in academic literature. We expect our approach to increase the coverage of the ICPSR Bibliography product by supporting the curation of resources for ingest into the ICPSR Bibliography. Our goal is to improve our ability to detect literature in which ICPSR datasets are informally referenced. We are prioritizing recall over precision because we first want to capture as many candidate references as possible to review and determine if they are true instances of data reuse.

We will make several changes to adapt our approach to the ICPSR Bibliography. First, we will incorporate available training data from the ICPSR Bibliography team, who maintain detailed notes providing evidence of data use for citations that are added to the ICPSR Bibliography. The notes will be matched against available full text publications to extend our training corpus.

Second, we will explore whether the usage patterns we detected in the competition corpus are applicable to resources in the ICPSR Bibliography. Given that data citations tend to be located in the methods section of articles [8, 11, 12], we were surprised that many of the data citations in the Coleridge corpus were detected elsewhere; we also noted an absence of known phrases and words indicating data usage in the training corpus. We will evaluate the extent of these differences when retraining our classifier and include only the most important features, which we expect will allow the model to generalize. As before, we will select a model based on recall criteria.

Finally, we will train the NER model to include ICPSR-specific instances of data citation. We will also evaluate our NER pipeline more extensively to establish an experimental baseline and detect issues like overfitting. In production, the model will support the ICPSR Bibliography team by surfacing and prioritizing candidate documents and sections of text within them for review. This research supports a broader effort to measure the impact of research data by increasing the coverage of the ICPSR Bibliography and the fidelity of data reuse metrics that rely on it.

[5] https://catalog.data.gov/dataset

References

[1] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[2] Martin Fenner, Daniella Lowenberg, Matt Jones, Paul Needham, Dave Vieglais, Stephen Abrams, Patricia Cruse, and John Chodacki. Code of practice for research data usage metrics release 1. Technical report, PeerJ Preprints, 2018. doi: 10.7287/peerj.preprints.26505v1.

[3] Libby Hemphill, Amy Pienta, Sara Lafia, Dharma Akmon, and David Bleckley. How do properties of data, their curation, and their funding relate to reuse? doi: 10.7302/1639, 2021.

[4] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020. doi: 10.5281/zenodo.1212303.

[5] Coleridge Institute. Rich context workshop summary report. Technical report, Coleridge Institute, 2019. Report at coleridgeinitiative.org/richcontext/richcontextworkshop/.

[6] Hyungrok Kim, Kinam Park, and Sae Hyong Park. Rich context competition: Extracting research context and dataset usage information from scientific publications. Report at https://rokrokss.com/assets/cv/rcc09.pdf.

[7] Brigitte Mathiak and Katarina Boland. Challenges in matching dataset citation strings to datasets in social science. *D-Lib Magazine*, 21(1/2):23–28, 2015. doi: 10.1045/january2015-mathiak.

[8] Christine Mayo, Todd J Vision, and Elizabeth A Hull. The location of the citation: changing practices in how publications cite original data in the dryad digital repository. *International Journal of Digital Curation*, 11(1):150–155, 2016. doi: 10.2218/ijdc.v11i1.400.

[9] Hailey Mooney. Citing data sources in the social sciences: do authors do it? *Learned Publishing*, 24(2):99–108, 2011. doi: 10.1087/20110204.

[10] Elizabeth Moss, Christin Cave, and Jared Lyle. Sharing and citing research data: A repository's perspective. In *American Casebook Series*. West Academic Publishing, 2015. hdl: 2027.42/115490.

[11] Elizabeth Moss and Jared Lyle. Opaque data citation: Actual citation practice and its implication for tracking data use, 2018. hdl: 2027.42/142393.

[12] Hyoungjoo Park, Sukjin You, and Dietmar Wolfram. Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. *Journal of the Association for Information Science and Technology*, 69(11):1346–1354, 2018. doi: 10.1002/asi.24049.

[13] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011. arXiv: 1201.0490.