Identifying data sharing and data reuse in full-text NIMH-funded papers

Poster No:

1250

Submission Type:

Abstract Submission

Authors:

Travis Riddle[1], Francisco Pereira[1], Adam Thomas[1]

Institutions:

[1]National Institute of Mental Health, Bethesda, MD

Introduction:

Identifying and measuring data sharing and data reuse serves a number of goals that are important for scientists, funding agencies, and the public more generally. Consequently, the unmet objective of an efficient and accurate system for identification and tracking of datasets is a conspicuous shortcoming of the larger open science community.

Methods:

The work we describe here uses natural language processing (NLP) and machine learning to identify data sharing and reuse statements in the full-text papers available on PubMed Central. We limited the scope of our investigation to just those papers that listed funding from the NIMH per Federal Reporter. We obtained the full text of 57,771 papers published after 2008. These papers were linked to 11,987 NIMH grants awarded to a total of 7,342 primary investigators. We then split these documents into sentences and labeled a subset of sentences with two types of labels:

- A sentence is considered an instance of "data sharing" if the authors are indicating that the data that they generated for the paper are deposited and available in a public repository of some sort.
- A sentence is considered "data reuse" if the authors are making reference to a specific shared dataset. Here we are defining reuse as broadly as possible. We did not attempt to confirm the data was used in any analysis, but only that the reference was to a specific dataset. Typically when there's a question, we err on the side of inclusion. Brain atlases and other types of shared data products are in this category.

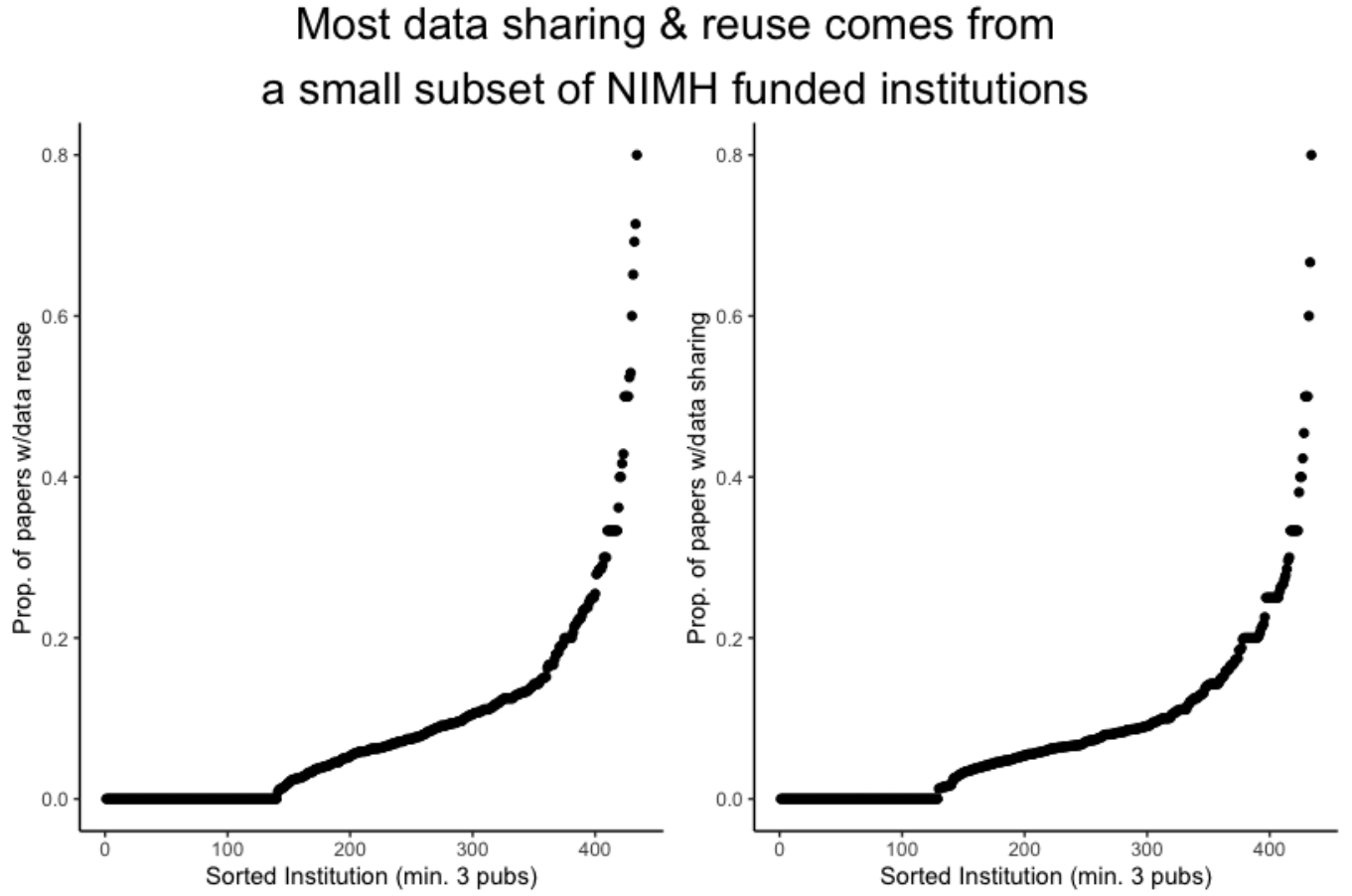
Because of the expected low base rates, we used regular expression matching of known data repositories, presence of a URL, lists published by repositories of papers that are known to have used their service, and active learning to maximize the likelihood of obtaining positive training examples. In total, we labeled 1,798 sentences for instances of data sharing, of which 71 were indicated as instances of data sharing. We labeled a partially overlapping set of 1,798 sentences for instances of data reuse, of which 129 were indicated as instances of data reuse.

We used standard NLP techniques to obtain input features to train an AdaBoost classifier to identify instances of data sharing and data reuse. Performance was evaluated using stratified 3-fold cross validation.

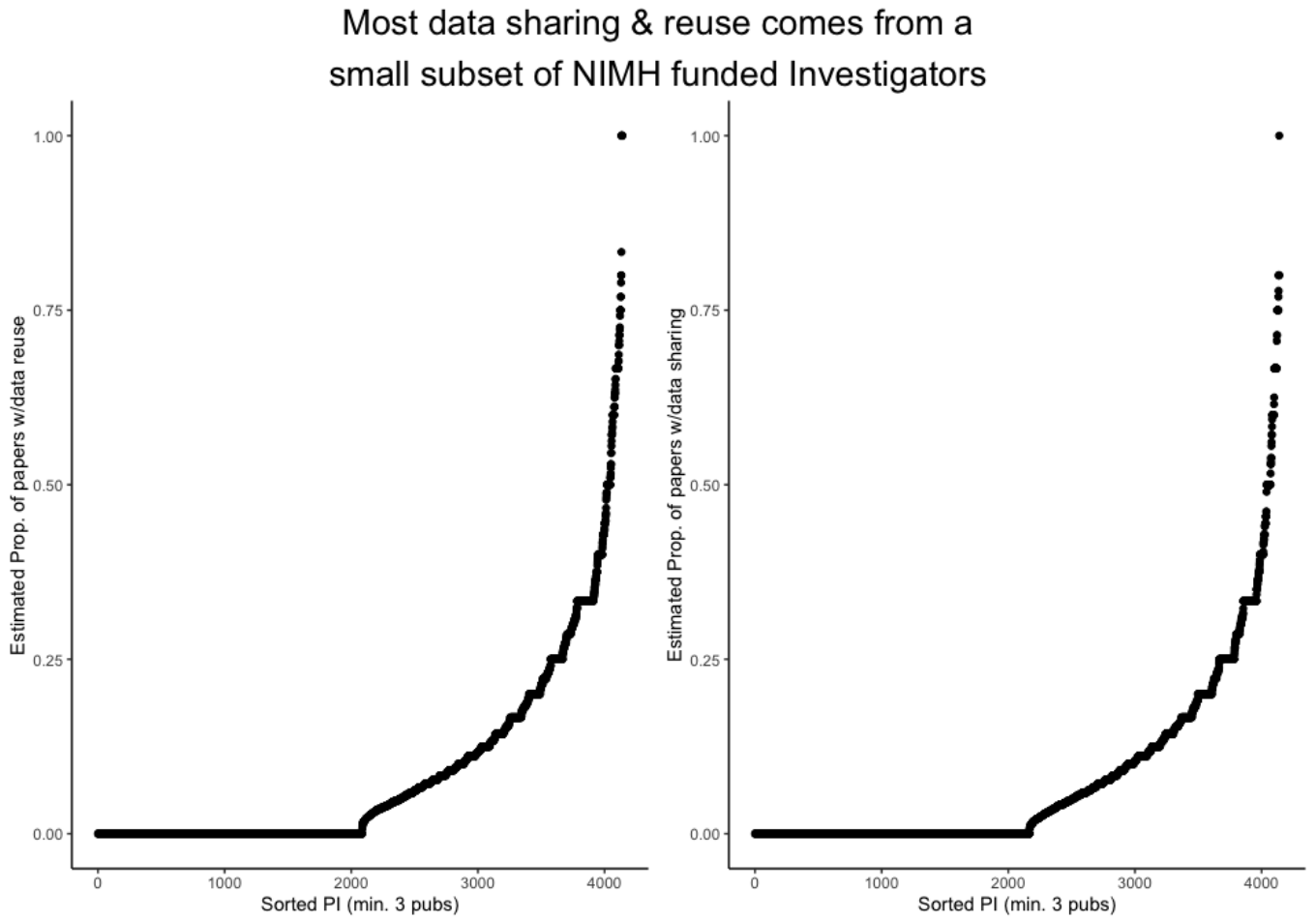
Results:

Generally, precision is higher than recall. If we average across folds and label types and extrapolate this performance, we expect our labels to accurately identify an instance of data sharing or reuse 70 percent of the time. Additionally, we expect to accurately identify (recall) 59 percent of instances of data sharing or reuse.

Accounting for these performance metrics, after retraining on the full labeled dataset, and extrapolating from our cross-validation performance, we expect that 4,214 (7.3%) papers contain instances of data sharing and 5,167 (8.9%) contain instances of data reuse, of which 1,777 and 2,179, respectively are expected to be incorrect predictions. Additionally, we are likely to miss 1,662, and 2,038 instances of data sharing and data reuse, respectively. These results indicate very low rates of data sharing and reuse. Of all 434 institutions that published at least 3 NIMH funded papers, we predicted just 18/25 have data sharing/reuse statements in more than 30% of their papers (Figure 1). Similarly, of all 4,139 PI's that published at least 3 NIMH funded papers, just 301/381 are expected to have data sharing/reuse statements in more than 30% of their papers (Figure 2). Our poster will feature an error analysis and an analysis of the features of papers labeled with and without data sharing and data reuse.



·Figure 1: Estimated Data sharing/reuse rates by institution



·Figure 2: Estimated Data sharing/reuse rates by investigator

Conclusions:

We anticipate that additional labeled data will help improve and stabilize performance of these methods. In the future, we also intend to explore the effectiveness of alternative approaches, including using a gold-standard list of dataset DOIs derived from Datacite.

Language:

Language Other

Modeling and Analysis Methods:

Classification and Predictive Modeling
Other Methods

Neuroinformatics and Data Sharing:

Databasing and Data Sharing [1]
Informatics Other [2]

Keywords:

Computing

Informatics
Language
Other - Natural Language Processing

[1|2]Indicates the priority used for review

My abstract is being submitted as a Software Demonstration.

No

Please indicate below if your study was a "resting state" or "task-activation" study.

Other

Healthy subjects only or patients (note that patient studies may also involve healthy subjects):

Healthy subjects

Are you Internal Review Board (IRB) certified? Please note: Failure to have IRB, if applicable will lead to automatic rejection of abstract.

No

Was any human subjects research approved by the relevant Institutional Review Board or ethics panel? NOTE: Any human subjects studies without IRB approval will be automatically rejected.

Not applicable

Was any animal research approved by the relevant IACUC or other animal research panel? NOTE: Any animal studies without IACUC approval will be automatically rejected.

Not applicable

Please indicate which methods were used in your research:

Computational modeling
Other, Please specify - Natural Language Processing

Provide references using author date format

Boland, K., Ritze, D., Eckert, K., & Mathiak, B. (2012). Identifying references to datasets in publications. In International Conference on Theory and Practice of Digital Libraries (pp. 150-161). Springer, Berlin, Heidelberg.

Duck, G., Nenadic, G., Brass, A., Robertson, D. L., & Stevens, R. (2013). bioNerDS: exploring bioinformatics' database and software use through literature mining. BMC bioinformatics, 14(1), 194.

Duck, G., Nenadic, G., Filannino, M., Brass, A., Robertson, D. L., & Stevens, R. (2016). A survey of bioinformatics database and software usage through mining the literature. PloS one, 11(6), e0157989.

Howison, J., & Bullard, J. (2016). Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. Journal of the Association for Information Science and Technology, 67(9), 2137-2155.

Lu, M., Bangalore, S., Cormode, G., Hadjieleftheriou, M., & Srivastava, D. (2012). A dataset search engine for the research document corpus. In 2012 IEEE 28th International Conference on Data Engineering (pp. 1237-1240). IEEE

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. PloS one, 2(3), e308.

van de Sandt, S., Nielsen, L. H., Ioannidis, A., Muench, A., Henneken, E., Accomazzi, A., ... & Dallmeier-Tiessen, S. (2019). Practice meets Principle: Tracking Software and Data Citations to Zenodo DOIs. arXiv preprint arXiv:1911.00295.