A Study of Computational Reproducibility using URLs Linking to Open Access Datasets and Software

Lamia Salsabil
Jian Wu
Muntabir Hasan Choudhury
Old Dominion University
Norfolk, Virginia, USA
lsals002@odu.edu,j1wu@odu.edu

William A. Ingram
Edward A. Fox
Virginia Polytechnic Institute and State University
Blacksburg, Virginia, USA
waingram@vt.edu,fox@vt.edu

Sarah M. Rajtmajer
C. Lee Giles
Pennsylvania State University
University Park, Pennsylvania, USA
smr48@psu.edu,clg20@psu.edu

ABSTRACT

Datasets and software packages are considered important resources that can be used for replicating computational experiments. With the advocacy of Open Science and the growing interest of investigating reproducibility of scientific claims, including URLs linking to publicly available datasets and software packages has become an institutionalized part of research publications. In this preliminary study, we investigated the disciplinary dependency and chronological trends of including open access datasets and software (OADS) in electronic theses and dissertations (ETDs), based on a hybrid classifier called OADSClassifier, consisting of a heuristic and a supervised learning model. The classifier achieves the best F1 of 0.92. We found that the inclusion of OADS-URLs exhibited a strong disciplinary dependence and the fraction of ETDs containing OADS-URLs has been gradually increasing over the past 20 years. We developed and share a ground truth corpus consisting of 500 manually labeled sentences containing URLs from scientific papers. The dataset and source code are available at https://github.com/lamps-lab/oadsclassifier.

CCS CONCEPTS

• **General and reference** → Empirical studies; • **Information systems** → *Digital libraries and archives*; • **Computing methodologies** → **Supervised learning**; *Supervised learning by classification*; **Information extraction**.

KEYWORDS

reproducibility, ETD, language model, open access

ACM Reference Format:
Lamia Salsabil, Jian Wu, Muntabir Hasan Choudhury, William A. Ingram, Edward A. Fox, Sarah M. Rajtmajer, and C. Lee Giles. 2022. A Study of Computational Reproducibility using URLs Linking to Open Access Datasets and Software. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3487553.3524658

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France.

ACM ISBN 978-1-4503-9130-6/22/04…$15.00
https://doi.org/10.1145/3487553.3524658

1 INTRODUCTION

Generally, reproducibility can be defined as *the ability for a researcher to duplicate the result of a prior study using the same materials as were used by the original investigators* [1, 8]. Results can be obtained using physical experiments–involving real-world equipment, objects and human subjects–or computational experiments. Since the inception of the Internet, there has been a growing number of research papers using computational methods to perform numerical simulations, or mine big data using machine learning and deep learning models [11, 15]. More and more papers include URLs linking to open access datasets and software (OADS) to make their work more transparent and easier to reproduce. Venues, in increasing numbers, encourage or require submitted papers to include URLs linking to OADS. Many OADS refer to standard training and testing corpora, e.g., ImageNet[1], or widely adopted software packages, e.g., BERT[2]. However, there are still a large number of OADS-URLs that are less well known, yet potentially useful for researchers. A method to automatically identify these URLs would facilitate building repositories supporting computational reproducibility studies in multiple disciplines.

Although recognizing URLs can be relatively straightforward using regular expressions, not all URLs link to OADS. Discovering URLs linking to OADS usually requires referring to the context around the target URL. For example, in Table 1, the context makes clear to readers that only the URL in the first sentence links to OADS. However, manually examining research papers to extract OADS is laborious and impractical given the rapid growth in the number of research papers [9], and there is no automation of this task, to best of our knowledge. To overcome this limitation, we propose a hybrid method to automatically identify OADS-URLs. We implemented this method in a pipeline and applied it to electronic theses and dissertations (ETDs).

This paper reports on the *disciplinary dependency and chronological trends of OADS-URLs identified in ETDs.* An ETD usually represents the major contribution of a student pursuing an academic degree. We have collected the full-text and metadata of about 500,000 ETDs published before 2021 [16] by crawling library repositories of universities in the United States. These ETDs cover both STEM and non-STEM disciplines. The relatively long documents, heterogeneous fields of study, and relatively broad span of years make this corpus ideal for our study.

[1] http://www.image-net.org
[2] https://github.com/google-research/bert

Table 1: Sentences containing OADS and non-OADS URLs.

Sentences containing URLs	Category
The secondary structures (alpha helices, beta strands and random coils) of the protein were predicted by using bioinformatics tools available on website http://npsa-pbil.ibcp.fr.	OADS
A supplementary appendix may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1111/puar.12797.	non-OADS

2 RELATED WORK

There has been growing interest in assessing and verifying the reproducibility of published findings, especially in social and behavioral sciences (SBS) [2, 4]. In a recent study, the authors attempted to identify important features that exhibited relatively strong correlation with experimental reproducibility in a corpus of SBS papers [18]. Other work has likewise tried to predict replicability of a corpus of SBS papers using a set of shallow features [19]. However, OADS-URLs were not included.

Computational reproducibility has been studied in several recent papers. One paper studied the URLs linking to datasets, focusing on papers produced by ACM SIGMOD and PVLDB [12]. The authors used a simple keyword-based method to search for links to source materials. Example keywords were "http", "online," etc. If the link was found active, they considered the resource to be available without distinguishing whether the URL truly linked to OADS. In another study, of social science papers, tf-idf and cosine similarity were used, but tf-idf is known to be less effective than word embedding models when representing semantic [7].

Färber et al. (2020) analyzed the quality and usage of GitHub code repositories using the Microsoft Academic Graph (MAG) [6]. The authors found a strong bias towards specific computer science areas (e.g., machine learning) and publication venues. The authors claimed that the set of URLs from MAG was more complete and precise than directly extracting URLs from full-text, but they did not provide details on the approaches used. In other work, authors studied 1.4 million Jupyter notebooks from GitHub, with the purpose of providing insights into the reproducibility of real notebooks [13]. They found that only 24.11% of notebooks executed without errors and only 4.03% produced the same results. URLs used in the above two studies were limited to GitHub links and therefore papers containing these URLs were published mostly after 2010.

Our work incorporates any URLs under HTTP or FTP protocols. We characterize the dependency and trends using the ETDs encompassing *multiple disciplines.*

3 CLASSIFYING OADS URLS.

3.1 Architecture Overview

A schematic architecture of our pipeline is depicted in Figure 1. The pipeline consists of the following modules.

(1) **PDF to text conversion.** First, PDFs of the papers were converted to text files. By comparing PDFMiner and PyPDF2, we found that a portion of text files converted by PyPDF2 removed white spaces between words, making it impossible to segment sentences. Therefore, PDFMiner was employed for conversion.

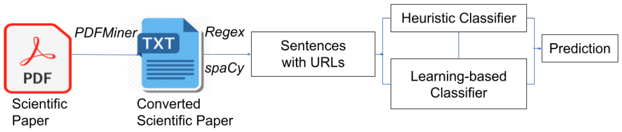

Figure 1: OADS URL classification pipeline

(2) **Sentence segmentation.** Next, we use SpaCy[3] for tokenizing the text into sentences. The SpaCy library was imported first, and then, to tokenize sentences, the English language model of SpaCy was loaded to iterate over the tokens of text.

(3) **Extraction of sentences with URLs.** We use the following regular expression to detect URLs in a sentence. Sentences containing URLs were then extracted.

```
(http|https|ftp|ftps)\:\/\/[a-zA-Z0-9\-\.]+\.[a-zA-Z]{2,3}(\/\S*)?
```

(4) **URL classification.** A hybrid method consisting of a heuristic model and a learning-based model was used to classify sentences containing URLs. Here, we assume that URLs contained in the same sentence have the same category. Our analysis indicated that out of 500 sentences, more than 93% of sentences contain only one URL, indicating that the number of URLs is roughly consistent with the number of sentences. For convenience, we refer to URLs linking to OADS as OADS-URLs and ETDs containing OADS-URLs as OADS-ETDs.

3.2 Heuristic Classifier

We observed that the majority of publisher URLs do not link to OADS. Therefore, we considered a simple heuristic method to exclude URLs that end with `.pdf` or link to publishers. We built a controlled list including 54 major publishers such as Springer, Wiley, and Sagepub. This heuristic method excludes non-OADS URLs with high accuracy, so they do not need to be classified by the learning-based model. However, we will investigate in our experiments whether a language model alone can achieve higher performance without "knowing" the URL's domains.

3.3 Learning-based Classifier

The learning-based model encodes a sentence using a pre-trained language model. We compare three transformer-based language models, namely, BERT [5], RoBERTa [10], and DistilBERT [14]. Because these three models were trained with general text, we also compare a document level embedding model, SPECTER [3], trained on academic documents. The maximum sequence length for BERT, DistilBERT, and RoBERTa was 512. The "bert-base-uncased", "roberta-base", and "distilbert-base-uncased" architectures were used for BERT, RoBERTa, and DistilBERT, respectively. To avoid overfitting, we tried different dropout values. The model performed well with a dropout rate 0.2. The output dimensions for BERT,

[3] https://spacy.io/

RoBERTa, DistilBERT, and SPECTER were 768. The vector representations were used to train and test a binary logistic regression (LR) classifier.

3.4 Hybrid Models

To effectively use labeled data and maximize the performance, we compared three hybrid models depending on whether the heuristic classifier was used in training and/or testing. The model with the highest F1 was adopted for our analysis.

(1) **No heuristic classifier.** In this model, all sentences in the training (testing) corpus were encoded into vectors and used for training (testing) the LR classifier.
(2) **Heuristic classifier for test data only.** The same as (1) except that the heuristic classifier was first applied to the testing data. The remaining sentences were classified using the LR classifier.
(3) **Heuristic classifier for training and test data.** The same as (1) except that the heuristic classifier was first applied to both training and testing data before using the LR classifier.

We also investigated whether the URLs provide useful information that improves sentence representation. To this end, we prepared two sets of sentences, one with original URLs masked with the word "URL" and the other with original URLs.

4 DATA

The ground truth dataset included 500 sentences containing URLs extracted from CORD-19 [17] and an in-house SBS paper corpus. The dataset was independently labeled as OADS and non-OADS by two graduate students, with an 83.6% consensus rate. A domain expert helped resolve differences. Several URLs were difficult to label because of the ambiguity of the sentences containing those URLs. For example, in the sentence "For more information, see: http://www.icpsr.umich.edu/icpsrweb/icpsr/studies/4607", there was little information in the context indicating whether the URL linked to OADS. In these cases, we visited the website the URL linked to. When labeling URLs, we focus on determining the nature of the contents. An OADS-URL may not necessarily be alive. The final ground truth contains 248 samples labeled as OADS, and the rest labeled as non-OADS. It was randomly split into 400 training samples and 100 test samples.

We randomly selected 100,000 ETDs from about 450k ETDs [16]. The entire dataset was collected by crawling 42 university libraries. A fraction of ETD metadata provided by the libraries was incomplete. Certain fields such as years were missing. All ETDs we selected contained values in the "year" and "department" fields.

Using PDFMiner, we converted 96,842 ETDs from PDF to text files. The metadata provided by the libraries contained over 60 departments. Because many departments were closely related, we consolidated departments into 18 disciplines (Figure 2) using the *Outline of Academic Disciplines* from Wikipedia[4].

Table 2: Precision (*P*), recall (*R*), and *F1*-scores for different hybrid models. The bold row has the highest F1.

Hybrid Model	Masking URLs			Original URLs		
	P	***R***	***F1***	***P***	***R***	***F1***
No heuristic classifier	0.86	0.72	0.81	0.86	0.89	0.89
Heuristic classifier for test data	0.86	0.90	0.89	0.87	0.95	0.91
Heuristic classifier for train and test data	**0.86**	**0.93**	**0.89**	**0.87**	**0.98**	**0.92**

Table 3: Precision (*P*), recall (*R*), and *F1*-scores of the OADSClassifier using different language models.

Language Model	Masking URLs			Original URLs		
	P	***R***	***F1***	***P***	***R***	***F1***
BERT	0.80	0.90	0.85	0.86	0.90	0.88
DistilBERT	**0.86**	**0.93**	**0.89**	**0.87**	**0.98**	**0.92**
RoBERTa	0.68	0.88	0.78	0.74	0.95	0.83
SPECTER	0.77	0.80	0.79	0.78	0.88	0.83

5 EXPERIMENTAL RESULTS

5.1 Hybrid Classifier Performance

We first compare the three hybrid models proposed in Section 3.4. The performance was evaluated using standard metrics: precision, recall, and F1-score. The results are tabulated in Table 2. Due to space constraints, we only show the performance with DistilBERT as the language model. The results indicated that adding the heuristic classifier for both training and testing data achieved the highest F1=92%. Masking URLs[5] decreases the performance by 3%.

Next, we investigate the effect of language models on the performance. Table 3 demonstrates that the best F1=0.92 was achieved using DistilBERT, leaving original URLs preserved in sentences. The BERT+LR model achieved the second best result with F1=0.88. Table 3 also shows that in general the classifier achieves a higher F1-score if URLs are not masked, indicating that URLs contain useful information that aids generating a better sentence representation. We attribute this to the WordPiece tokenizer that was used in BERT and its variants. Although an arbitrary URL is likely to be an out-of-vocabulary token, the URL can be further parsed into subword tokens. Certain subword tokens, such as the ones comprising words like "data" and "software", could be features of OADS-URLs.

5.2 Disciplinary Dependency

By applying the OADSClassifier to the ETDs we selected, we identified 51,201 (∼ 14%) sentences containing OADS-URLs out of 369,802 sentences containing URLs. The identified OADS-URLs appear in 15,951 ETDs, i.e., about 16.3% of the ETDs in our corpus.

[4] https://en.wikipedia.org/wiki/Outline_of_academic_disciplines

[5] i.e., replacing a URL https://foo.com/bar with a string "URL".

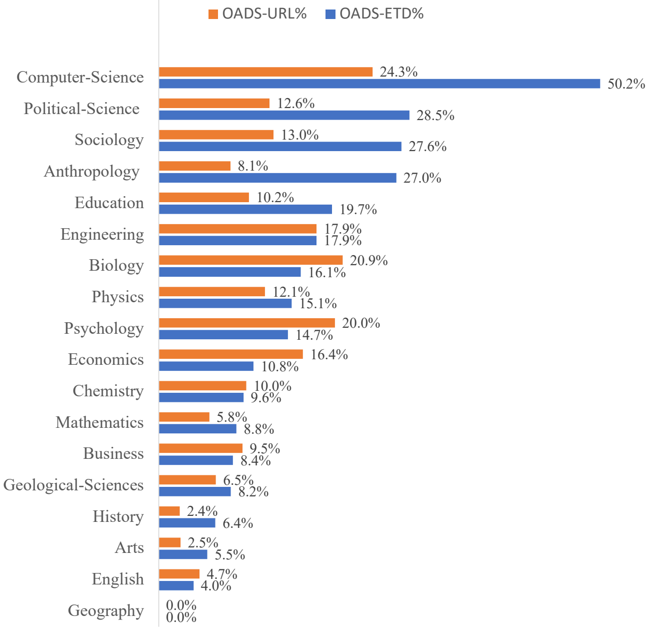

Figure 2: Dependency of the fractions of OADS-URL and OADS-ETD for academic disciplines.

Next, we study how the inclusion of OADS-URLs changes depending on academic discipline. Figure 2 shows two fractions:

$$\text{OADS-ETD\%} = \frac{N_{\text{OADS–ETD}}}{N_{\text{ETD}}}, \quad \text{OADS-URL\%} = \frac{N_{\text{OADS-URL}}}{N_{\text{URL}}}. \tag{1}$$

For a given discipline, $N_{\text{OADS–ETD}}$ is the number of ETDs containing OADS-URLs and N_{ETD} is the total number of ETDs in that discipline. Similarly, $N_{\text{OADS–URL}}$ is the number of OADS-URLs and N_{URL} is the total number of URLs in that discipline. Figure 2 shows several interesting results. (1) Computer Science has the highest fraction of OADS-ETD% (50.2%), consistent with Figure 7 by Färber et al. [6], which indicates most computer science ETDs include OADS-URLs. (2) ETDs in social sciences (e.g., Political science, Sociology, Anthropology, and Education) contain a relatively higher fraction of OADS-ETD% than STEM disciplines (e.g., Engineering, Biology, and Physics). In particular, we did not find any of the 717 Geography ETDs containing OADS-URLs. (3) Certain disciplines have a very small fraction of OADS-ETDs ($<$ 10%), such as Chemistry (9.6%), Business (8.4%), and Geological-Sciences (8.2%), indicating that it is less frequent to find computationally reproducible works in these disciplines. (4) The OADS-URL% exhibits a relatively even distribution. Computer Science has the highest OADS-URLs% (24.3%), followed by Biology (20.9%) and Psychology (20.0%), indicating that most URLs in ETDs in these fields ($>$ 75%) do not link to OADS. This phenomenon is more prominent for disciplines such as Chemistry (10%), Mathematics (5.8%), and Geological-Sciences (6.5%).

5.3 Chronological Trends

In the context of our sample of ETDs, we analyzed the chronological trends of OADS-URLs. We acknowledge that our sample may be biased, since there are more recent ETDs embargoed (so not present in the dataset) than older ETDs, and the fraction embargoed is discipline specific. Further, due to our collection process, the number of ETDs in the period starting 2019 is less than what would be expected if that process were repeated in a few years, and the numbers before 2010 are likewise low; these factors could lead to uncertainties in the analysis. Nevertheless, we hypothesize, for this preliminary analysis, that these issues do not significantly influence our findings.

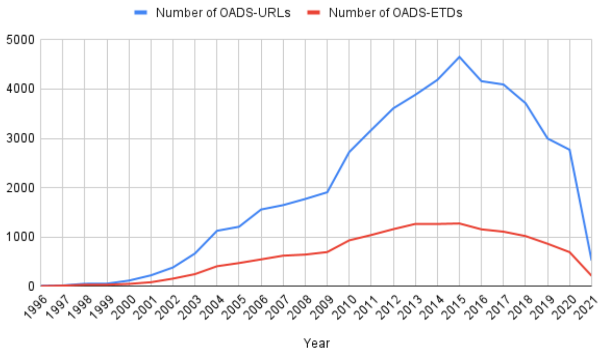

Figure 3: Numbers of OADS-URLs and ETDs containing OADS URLs as a function of publication year in our dataset.

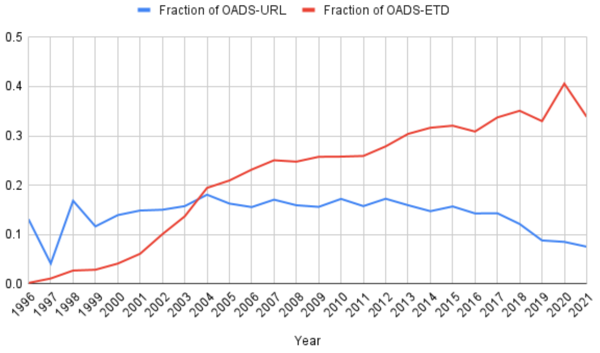

Figure 4: Fractions of OADS-URLs (blue) and ETDs containing OADS-URLs (red) as a function of year.

Figure 4 illustrates the fraction of OADS-URLs and the fraction of OADS-ETDs as a function of year; we use the definitions given in Eq. (1). First, the fraction of OADS-ETDs has been gradually increasing from less than 5% in 2000 to more than 25% in 2010 to about 40% in 2020. Second, the fraction of OADS-URLs seemed relatively stable after year 2000. Since 2016, this fraction has gradually decreased from 15% to about 10% in 2019–2020. There are two possible reasons that could contribute to this trend. (a) The growth of non-OADS URLs in ETDs, and (b) the selection bias (as seen in Figure 3) due to a weak correlation between embargoed ETDs and the inclusion of OADS-URLs. Further investigations are needed regarding (a) and (b).

6 CONCLUSIONS AND DISCUSSION

Using OADS-URLs as a proxy, we studied the computational reproducibility of academic documents, focusing on ETDs collected from universities in the USA. One key contribution is a model that automatically identifies sentences containing OADS-URLs from research papers. This model achieved the best F1 of 0.92. Our analysis for URLs in ETDs found that the inclusion of OADS-URLs exhibited a strong dependency on disciplines. The fraction of OADS-ETDs gradually increases over the past 20 years. The fraction of OADS-URLs was relatively stable between 2000 and 2015.

This work has the following limitations. First, the training and evaluation were based on samples drawn from CORD-19 and SBS papers; we assumed the model could be transferred to other academic disciplines. The results in Table 3 indicate that the language model trained on general text (i.e., DistilBERT) beats the language model trained on academic document (i.e., SPECTER), indicating that the language discrepancy between disciplines may not be significant and thus the model could be transferred for this task. Second, a more complete sample is needed to reveal more accurate dependencies and trends after 2016. In addition to addressing the above limitations, the future plans include developing a multi-class classifier that distinguishes whether a OADS-URL links to a dataset or to software, and whether they were published by the authors or by a third party.

ACKNOWLEDGMENTS

This work was partially supported by the Institute of Museum and Library Services through grant LG-37-19-0078-19, and by the Defense Advanced Research Projects Agency (DARPA) under cooperative agreement No. W911NF-19-2- 0272. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

REFERENCES

[1] John T Cacioppo, Robert M Kaplan, Jon A Krosnick, James L Olds, and Heather Dean. 2015. Social, behavioral, and economic sciences perspectives on robust and reliable science. *Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences* 1 (2015).

[2] Colin F. Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers, and Hang Wu. 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour* 2, 9 (2018), 637–644. https://doi.org/10.1038/s41562-018-0399-z

[3] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 2270–2282. https://doi.org/10.18653/v1/2020.acl-main.207

[4] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015). https://doi.org/10.1126/science.aac4716 arXiv:https://science.sciencemag.org/content/349/6251/aac4716.full.pdf

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. 4171–4186. https://aclweb.org/anthology/papers/N/N19/N19-1423/

[6] Michael Färber. 2020. *Analyzing the GitHub Repositories of Research Papers.* Association for Computing Machinery, New York, NY, USA, 491–492. https://doi.org/10.1145/3383583.3398578

[7] Behnam Ghavimi, Philipp Mayr, Sahar Vahdati, and Christoph Lange. 2016. Identifying and improving dataset references in social sciences full texts. *arXiv preprint arXiv:1603.01774* (2016).

[8] Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. 2016. What does research reproducibility mean? *Science Translational Medicine* 8, 341 (2016), 341ps12–341ps12. https://doi.org/10.1126/scitranslmed.aaf5027 arXiv:https://www.science.org/doi/pdf/10.1126/scitranslmed.aaf5027

[9] Peder Larsen and Markus Von Ins. 2010. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics* 84, 3 (2010), 575–603.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints* (July 2019). arXiv:1907.11692 [cs.CL]

[11] Sheshera Mysore, Zachary Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, and Elsa Olivetti. 2019. The Materials Science Procedural Text Corpus: Annotating Materials Synthesis Procedures with Shallow Semantic Structures. In *Proceedings of the 13th Linguistic Annotation Workshop.* Association for Computational Linguistics, Florence, Italy, 56–64. https://doi.org/10.18653/v1/W19-4007

[12] Mateusz Pawlik, Thomas Hütter, Daniel Kocher, Willi Mann, and Nikolaus Augsten. 2019. A Link is not Enough–Reproducibility of Data. *Datenbank-Spektrum* 19, 2 (2019), 107–115.

[13] João Felipe Pimentel, Leonardo Murta, Vanessa Braganholo, and Juliana Freire. 2019. A Large-Scale Study About Quality and Reproducibility of Jupyter Notebooks. In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*. 507–517. https://doi.org/10.1109/MSR.2019.00077

[14] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR* (2019). arXiv:1910.01108

[15] Sree Sai Teja Lanka, Sarah Rajtmajer, Jian Wu, and C. Lee Giles. 2021. *Extraction and Evaluation of Statistical Information from Social and Behavioral Science Papers.* Association for Computing Machinery, New York, NY, USA, 426–430. https://doi.org/10.1145/3442442.3451363

[16] Sami Uddin, Bipasha Banerjee, Jian Wu, William A Ingram, and Edward A Fox. 2021. Building A Large Collection of Multi-domain Electronic Theses and Dissertations. In *2021 IEEE International Conference on Big Data (Big Data).* IEEE Computer Society, 6043–6045. DOI:10.1109/BigData52589.2021.9672058

[17] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The Covid-19 Open Research Dataset. *CoRR* (2020). arXiv:2004.10706

[18] Jian Wu, Rajal Nivargi, Sree Sai Teja Lanka, Arjun Manoj Menon, Sai Ajay Modukuri, Nishanth Nakshatri, Xin Wei, Zhuoer Wang, James Caverlee, Sarah Michele Rajtmajer, and C. Lee Giles. 2021. Predicting the Reproducibility of Social and Behavioral Science Papers Using Supervised Learning Models. *CoRR* (2021). arXiv:2104.04580

[19] Yang Yang, Wu Youyou, and Brian Uzzi. 2020. Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proceedings of the National Academy of Sciences* 117, 20 (2020), 10762–10768.