SHORT: Can citations tell us about a paper's reproducibility? A case study of machine learning papers

Rochana R. Obadage
Old Dominion University
Norfolk, VA, USA
oruma001@odu.edu

Sarah M. Rajtmajer
IST, Pennsylvania State University
University Park, PA, USA
smr48@psu.edu

Jian Wu
Old Dominion University
Norfolk, VA, USA
j1wu@odu.edu

ABSTRACT

The iterative character of work in machine learning (ML) and artificial intelligence (AI) and reliance on comparisons against benchmark datasets emphasize the importance of reproducibility in that literature. Yet, resource constraints and inadequate documentation can make running replications particularly challenging. Our work explores the potential of using downstream citation contexts as a signal of reproducibility. We introduce a sentiment analysis framework applied to citation contexts from papers involved in Machine Learning Reproducibility Challenges in order to interpret the positive or negative outcomes of reproduction attempts. Our contributions include training classifiers for reproducibility-related contexts and sentiment analysis, and exploring correlations between citation context sentiment and reproducibility scores. Study data, software, and an artifact appendix are publicly available at *https://github.com/lamps-lab/ccair-ai-reproducibility*.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Information systems** → **Data analytics**.

KEYWORDS

Citation Contexts, Reproducibility, Machine Learning, Sentiment Analysis, Science of Science

ACM Reference Format:
Rochana R. Obadage, Sarah M. Rajtmajer, and Jian Wu. 2024. SHORT: Can citations tell us about a paper's reproducibility? A case study of machine learning papers. In *Proceedings of ACM Conference on Reproducibility and Replicability (ACM REP'24).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In the rapidly advancing fields of machine learning (ML) and artificial intelligence (AI), emerging technologies are often advancements, refinements, or assemblies of existing ones. Because of this, reproducibility is central to progress in these fields. Yet, the growing complexity of ML research, hardware and software resource constraints, code and data with insufficient documentation, proprietary datasets, and current incentive structures have made the direct reproduction of existing models and findings infeasible in many cases. Indeed, the ML/AI community is beginning to demonstrate that it too faces a crisis of reproducibility [4, 19]. Directly reproducing reported results is the most reliable method to test the reproducibility of a paper, but this method does not scale to millions of research papers. Existing efforts to automatically assess reproducibility and replicability either use shallow features directly extracted from paper content, such as bibliographic, statistical, and semantic features [27], or latent features such as word embeddings [28]. As yet, no automatic approach to reproducibility or replicability assessment has demonstrated good enough performance to be useful in real-world scenarios.

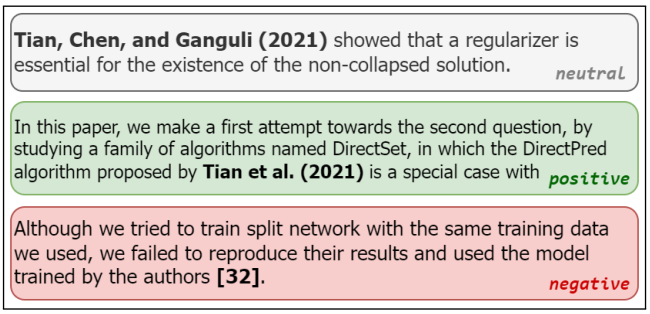

Figure 1: Examples of citation context with different reproducibility sentiments.

Our work explores opportunities to mine the text around downstream citations of a paper–**citation context** (Figure 1)–for cues indicating that an author has successfully or unsuccessfully replicated another paper's work in the course of their own. This is motivated by an understanding that most reproductions and replications are relatively informal (vs. explicitly framed as a replication study) and occur commonly in ML/AI in service to model comparisons against benchmarks, use of one technology in service to a different research aim, or similar. The reproducing or replicating author will note this in their own work, information which–if reliably extracted–could be a rich resource for understanding reproducibility and replicability in the field. Although prior research has performed classification on citation contexts for various purposes [11, 13, 15, 24], whether citation context can be used as a signal of reproducibility has not yet been studied. In this paper, we address this question by exploring the correlation between reproducibility scores and reproducibility sentiment of citation context. [1]

[1]We adopt the definition of reproducibility from [16] where a finding is deemed reproducible if consistent results are obtained using the same input data, computational

ACM REP'24, 978-1-4503-XXXX-X/18/06

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

Specifically, we train ML models to classify citation contexts based on their reproducibility-based sentiment. This approach falls under the umbrella of aspect-based sentiment analysis, a text analysis technique that classifies text by sentiment related to a given aspect [17]. Reproducibility scores are calculated based on the reports of direct reproducibility studies. We then study whether the reproducibility score is correlated with the normalized citation context count of a certain sentiment. Our contributions are as follows:

(1) We observe a correlation between reproducibility scores and citation context sentiment in a pilot dataset containing 41,244 citation contexts extracted from 130 ML papers.
(2) We build the first ground truth dataset of direct reproducibility scores of 22 ML papers and 1937 citation contexts, manually labeled into three reproducibility sentiment categories.
(3) We train and validate two ML models to classify citation context into reproducibility sentiments and obtain F1-scores ranging from 0.70 to 0.86.

2 RELATED WORK

Our study is related to prior work on reproducibility of ML and AI. While Gundersen et al. [10] discuss state-of-the-art of reproducibility in AI, Raff [19] evaluates the reproducibility of 255 papers published between 1984 to 2017, emphasizing the importance of factors beyond code availability. Akella et al. [3] discuss factors that contribute to reproducibility of ML findings and propose solutions.

We build upon work in sentiment analysis. Yousif et al. [29] propose a multitask learning model based on convolutional and recurrent neural networks that perform the citation sentiment and purpose classification. HuggingFace [26] contains a repository of open-source ML models for sentiment analysis tasks trained on various datasets. We used both supervised and unsupervised models trained on different datasets including tweets, social media posts, and citation contexts.

In terms of citation context classification, Cohan et al. [7] propose "structural scaffolds", a multitask model incorporating structural information of scientific papersfor classification of citation intent. Te et al. [24] compare methods for classification of critical vs. non-critical citation contexts. Budi et al. discuss citation meaning using sentiment, role, and citation function classifications [5].

3 DATASET

The dataset contains three types of scientific documents: **original studies** - original research papers/findings which serve as targets for a reproduction; **reproducibility studies (rep-studies)** - papers reporting attempts to reproduce a particular paper/finding; and, **citing papers** - papers that cite the original studies. We prepared our dataset in five steps (see Figure 2).

(1) Collect reproducibility studies from existing data sources
(2) Collect metadata for original studies
(3) Calculate the reproducibility score for each rep-study
(4) Collect citation contexts from citing papers
(5) Label citation contexts by reproducibility sentiments

steps, methods, and code, and conditions of analysis. This definition is consistent with the definitions of reproducibility adopted by ACM [1].

3.1 Reproducibility Studies

We collected the metadata for 145 rep-studies from existing data sources listed in Table 1. The Machine Learning Reproducibility Challenges (MLRC) contains 129 rep-studies of 114 papers (some rep-studies were conducted on the same original paper). Because the majority of papers were successfully reproduced, we supplement the data with 16 rep-studies by Ajayi et al. [2] including 5 successful and 11 unsuccessful rep-studies. The same input data, computational steps, methods, and analysis conditions were adopted in all rep-studies except that the experiments were conducted by different teams. Using the DOIs we collected metadata for all the 130 original studies from the Semantic Scholar Graph API (S2GA; [14]).

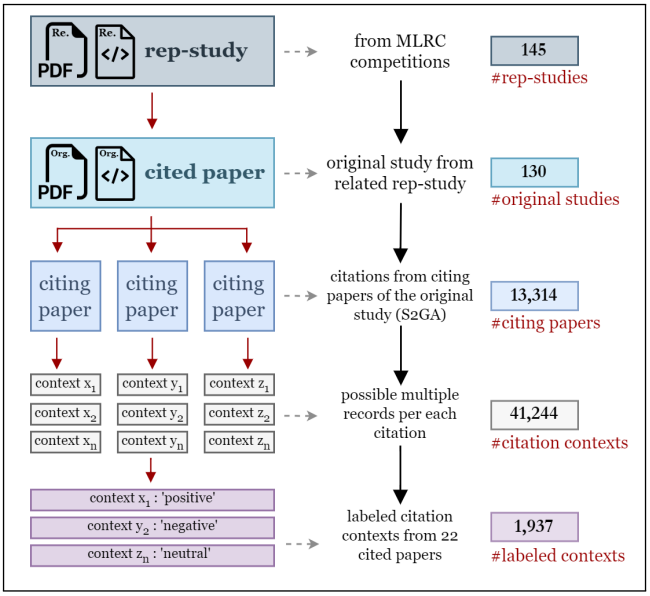

Figure 2: A schematic illustration of the data reduction and processing workflow.

3.2 Reproducibility Score Calculation

Traditionally, reproducibility scores are defined as dichotomous values, indicating reproducibility of a paper based on its primary finding. Here, we introduce an extended reproducibility score (rs_score) for a paper with *multiple* findings, given by the equation below. To calculate it, we perused full text of replication studies. The rs_score distribution is shown in Table 2.

$$\textbf{rs\_score} = \frac{\text{\# of successfully reproduced findings}}{\text{\# of total findings selected to reproduce}}$$

3.3 Citation Context Collection

We collected all citation contexts available for each of 130 original studies from S2GA [14] (Table 1). We collected 13,314 citations and 41,244 citation contexts. On average, each original paper was cited more than 3 times per citing paper.

Table 1: Data sources for selected reproducibility studies with the year of reproduction, the numbers of rep-studies and citation contexts for each data source.

Data Source	Year	#Rep-Studies	#Contexts
ICLR [6, 23]	2019	4	2102
NeurIPS [6, 23]	2019	10	9908
MLRC [6, 23]	2020	23	10798
MLRC [6, 23]	2021	47	6958
MLRC [6, 23]	2022	45	9869
TSR [2]	2023	16	1609
Total		145	41244

Table 2: Distribution of original papers and citation contexts over rs_scores. The rows corresponding to zero original papers (and citation context) are not shown. N_{pos} and N_{pos} for models M6 and M7 are defined in Section 5.2.

rs_score	#paper	#Context	M6		M7	
			N_{pos}	N_{neg}	N_{pos}	N_{neg}
0.0	11	1288	337	98	181	72
0.2	1	49	17	6	2	4
0.4	2	152	46	6	32	2
0.5	12	1967	669	147	469	114
0.6	5	284	90	27	45	10
0.7	4	140	33	9	26	2
0.8	3	161	31	9	29	6
1.0	89	37203	14521	2064	9516	1729

3.4 Building the Ground Truth

As a pilot study, we randomly selected 22 original papers (Table 1) and 1937 citation contexts. We manually labeled these citation contexts into 3 reproducibility sentiments:

- **positive**: the context hints about reproducibility (such as re-usage about the cited paper's data/code or the concept);
- **negative**: the context hints about irreproducibility (such as unavailability of the cited paper's data/code or unsuccessful attempts in reproducing);
- **neutral**: the context simply mentions (cites) the cited paper without any hints about reproducibility.

The ground truth dataset contains 158 positive (8.1%), 23 negative (1.2%), and 1756 neutral (89.7%) citation contexts. As the distribution is skewed, we down-sampled positive and neutral citation contexts to match the number of negative citation contexts, for a balanced ground truth subset containing 69 labeled citation contexts.

4 SENTIMENT ANALYSIS

We first performed aspect-based sentiment analysis by classifying citation context into the three sentiments above. To our knowledge, there are no ready-to-use models for this task, so we trained our own ML models using the balanced ground truth subset (Section 3.4). For comparison, we selected five pre-trained open-source multiclass sentiment analysis models from *HuggingFace* [26] (Table 3) based on the popularity. M1-M5 were trained/fine-tuned on social media posts, tweets, or generic datasets.

Table 3: Comparison of mean weighted average precision, recall, and F1-scores for M1-M5.

Model	Domain	mAP	mAR	mAF_1
M1 [12]	social media posts	0.46	0.43	0.34
M2 [22]	generic dataset	0.67	0.42	0.37
M3 [18]	tweets	0.63	0.51	0.48
M4 [8]	generic dataset	0.54	0.42	0.35
M5 [25]	generic dataset	0.39	0.48	0.41

Table 4: 5-fold cross-validation results for M6 and M7.

Model	Data	mAP	mAR	mAF_1
M6 (sentiment 3-class)	93	0.81	0.71	0.70
M7.1 (related/not related)	362	0.83	0.82	0.82
M7.2 (positive/negative)	46	0.92	0.83	0.86

We trained two in-house models using our data. The first (**M6**) leverages DistilBERT [21] fine-tuned using our ground truth data. Compared with BERT [9], DistilBERT is lighter, faster, and achieved the top performance for our previous reproducibility-related study [20]. Our second model is a hierarchical classifier (**M7**, combining M7.1 and M7.2) to verify our results obtained from M6. In the first step, we trained a binary classifier (M7.1) which classifies citation contexts as related to or not related to reproducibility. We used the full ground truth set (1937 labeled citation contexts) and merged positive and negative labels into a category called ***related***; neutral labels were categorized as ***not related***. In the second step, we fine-tuned a DistilBERT binary classifier (Table 3: *M7.2*) to classify citation contexts labeled as related as either ***positive*** or ***negative***.

5 RESULTS

5.1 Sentiment Analysis

The evaluation results of the five baseline methods are shown in Table 3. These models are evaluated on the balanced ground truth subset consisting of an equal number of positive, negative, and neutral citation contexts. These baseline models were not trained on our data, which explains why they do not perform well. To evaluate M6 and M7, we performed 5-fold cross-validation (Table 4) using the ground truth set supplemented with additional positive and neutral samples. M6 and M7 achieve F1-scores of 0.70 and M7 achieved an F1-score of 0.86, respectively. We note that we were not able to test M6 and M7 on the identical data as M1-M5 because M6 and M7 incorporate some of this data for training. Nevertheless we observe M6 and M7 perform significantly better than baselines and achieve reasonably good performance.

Next, M6 and M7 were applied to all 41,244 citation contexts. Applying M6 resulted in 15,744 positive (38.17%), 2366 negative (5.74%), and 23,134 neutral (56.09%) citation contexts. Applying

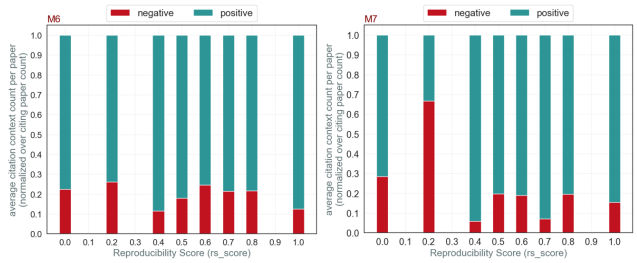

Figure 3: Normalized citation context sentiment counts vs. reproducibility scores using M6 (left) and M7 (right).

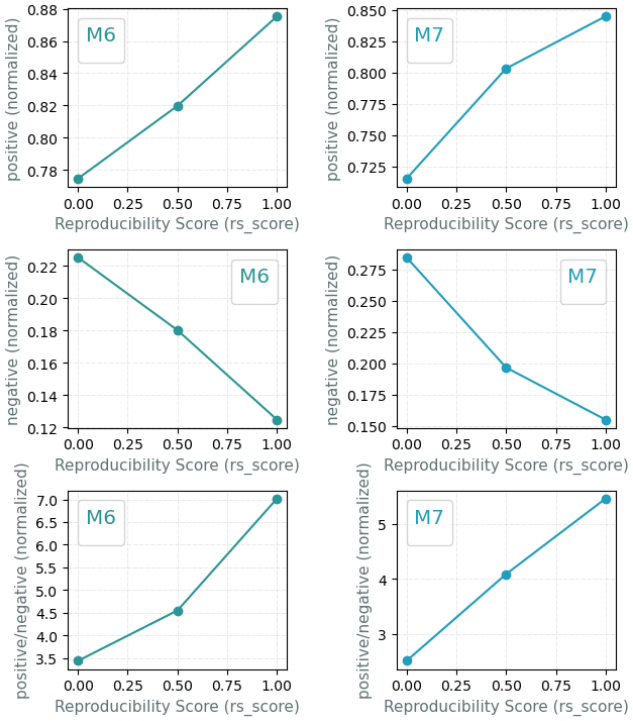

Figure 4: Normalized positive and negative citation context counts vs. rs_scores using M6 (left) and M7 (right).

M7 resulted in 10,300 positive (24.97%), 1939 negative (4.70%), and 29,005 (70.33%) neutral citation contexts.

5.2 Citation Context Sentiments vs. Reproducibility Scores

Our goal is to investigate the correlation between the reproducibility sentiment of citation contexts and the reproducibility scores of original papers. Because of the strongly skewed distribution of reproducibility scores (Table 2), simply using the numbers of citation contexts will not result in meaningful conclusions. Therefore, we normalized the number of citation contexts by a factor $Z = N_{\text{pos}}+N_{\text{neg}}$, in which N_{pos} and N_{neg} are the numbers of citation contexts labeled as positive and negative reproducibility-based sentiments, respectively. Therefore, the **normalized citation context counts**, i.e., the fraction of positive or negative citation contexts, are given by:

$$N'_{\text{pos}} = N_{\text{pos}}/Z, \quad N'_{\text{neg}} = N_{\text{neg}}/Z, \quad N'_{\text{pos}} + N'_{\text{neg}} = 1.$$

Figure 3 depicts N'_{pos} and N'_{neg} using M6 and M7.

We note that N'_{pos} or N'_{neg} at certain rs_scores are calculated based on a small number of papers/contexts. For example, there is only one paper whose rs_score is 0.2 and M7 only predicts 2 positive and 4 negative citation contexts (Table 2). The low citation counts in combination with the uncertainty introduced by the sentiment classification models may lead to large uncertainties of N'_{pos} and N'_{neg}. To obtain statistically meaningful results, we remove data points calculated based on less than 50 negative citation contexts, leaving three data points at rs_score = 0, 0.5, and 1.

Correlations between the normalized citation context count of positive or negative sentiment and the rs_scores, for M6 and M7, are shown in Figure 4. Given a cited paper, the fraction of positive citation contexts (N'_{pos}) increases with the reproducibility scores, and the fraction of negative citation contexts (N'_{neg}) decreases with the reproducibility scores. The ratio $r = N'_{\text{pos}}/N'_{\text{neg}}$ exhibits a magnified correlation with rs_score, with r ranging from about 3.5 to 7 for M6 and from 2.5 to 5.5 for M7. Because there are only three data points for each diagram, we did not calculate the correlation and regression coefficients.

6 DISCUSSION AND CONCLUSION

In this pilot study, we explored correlations between reproducibility-based sentiments of citation context and reproducibility scores using a total of 41,244 citation contexts. We trained two sentiment analysis models and achieved F1-scores of 0.70–0.86. Both models exhibited an increasing fraction of positive sentiment citation context with rs_score and a decreasing fraction of negative sentiment citation context with rs_score. The correlation is stronger in the ratio diagrams.

If our findings are verified using larger datasets, it suggests that it is possible to *statistically* estimate the reproducibility of ML papers using downstream citation contexts. More precisely, our work suggests that downstream mentions of a paper contain signals about the efforts ML researchers routinely undertake to reproduce one another's models and findings, often for purposes of extension or comparison, but which are not systematically reported as reproducibility studies. Nevertheless, this study does not imply to use of citation context sentiments to replace direct experiments to assess reproducibility. Rather, they may be useful as a surrogate to study the trends of reproducibility and its correlations with other factors for large corpora of ML papers when direct reproducibility studies are not feasible.

One limitation is the relatively low number of training data. In the future, we will extend our labeling to more papers with direct reproducibility studies, such as ML papers with ACM badges and papers with partial reproducibility scores ($0 < rs\_score < 1$) to verify and confirm our observations. Another potential limitation is the selection bias. Most rep-studies we adopted are from MLRC, which intentionally reproduces papers published in top-tier venues. This bias can be mitigated by collecting more rep-studies based on a homogeneous selection of venues.

REFERENCES

[1] ACM. 2020. . https://www.acm.org/publications/policies/artifact-review-badging Accessed March 28, 2024.

[2] Kehinde Ajayi, Muntabir Hasan Choudhury, Sarah M. Rajtmajer, and Jian Wu. 2023. A Study on Reproducibility and Replicability of Table Structure Recognition Methods. In *Document Analysis and Recognition - ICDAR 2023*, Gernot A. Fink, Rajiv Jain, Koichi Kise, and Richard Zanibbi (Eds.). Springer Nature Switzerland, Cham, 3–19.

[3] A. Akella, D. Koop, and H. Alhoori. 2023. Laying Foundations to Quantify the "Effort of Reproducibility". In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE Computer Society, Los Alamitos, CA, USA, 56–60. https://doi.org/10.1109/JCDL57899.2023.00018

[4] Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A Systematic Review of Reproducibility Research in Natural Language Processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 381–393. https://doi.org/10.18653/v1/2021.eacl-main.29

[5] Indra Budi and Yaniasih Yaniasih. 2023. Understanding the meanings of citations using sentiment, role, and citation function classifications. *Scientometrics* 128, 1 (01 Jan 2023), 735–759. https://doi.org/10.1007/s11192-022-04567-4

[6] RESCIENCE C. 2023. ReScience C — rescience.github.io. https://rescience.github.io/read/. [Accessed 10-02-2024].

[7] Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural Scaffolds for Citation Intent Classification in Scientific Publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 3586–3596. https://doi.org/10.18653/v1/N19-1361

[8] Souvick Das. 2022. Souvikcmsa/BERT-sentiment-analysis Hugging Face — huggingface.co. https://huggingface.co/Souvikcmsa/BERT_sentiment_analysis. [Accessed 10-02-2024].

[9] 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/

[10] Odd Erik Gundersen and Sigbjørn Kjensmo. 2018. State of the Art: Reproducibility in Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018). https://doi.org/10.1609/aaai.v32i1.11503

[11] Priyanshi Gupta, Yash Kumar Atri, Apurva Nagvenkar, Sourish Dasgupta, and Tanmoy Chakraborty. 2023. Inline Citation Classification Using Peripheral Context and Time-Evolving Augmentation. In *Advances in Knowledge Discovery and Data Mining*, Hisashi Kashima, Tsuyoshi Ide, and Wen-Chih Peng (Eds.). Springer Nature Switzerland, Cham, 3–14.

[12] Jochen Hartmann, Mark Heitmann, Christina Schamp, and Oded Netzer. 2021. The Power of Brand Selfies. *Journal of Marketing Research* (2021).

[13] MYRIAM HERNÁNDEZ-ALVAREZ, JOSÉ M. GOMEZ SORIANO, and PATRICIO MARTÍNEZ-BARCO. 2017. Citation function, polarity and influence classification. *Natural Language Engineering* 23, 4 (2017), 561–588. https://doi.org/10.1017/S1351324916000346

[14] Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul L Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. 2023. The Semantic Scholar Open Data Platform. *ArXiv* abs/2301.10140 (2023). https://api.semanticscholar.org/CorpusID:256194545

[15] Suchetha N. Kunnath, Drahomira Herrmannova, David Pride, and Petr Knoth. 2021. A meta-analysis of semantic classification of citations. *Quantitative Science Studies* 2, 4 (12 2021), 1170–1215. https://doi.org/10.1162/qss_a_00159 arXiv:https://direct.mit.edu/qss/article-pdf/2/4/1170/2007871/qss_a_00159.pdf

[16] Engineering National Academies of Sciences, Medicine, et al. 2019. Reproducibility and replicability in science. (2019).

[17] Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2022. Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey. *IEEE Transactions on Affective Computing* 13, 2 (2022), 845–863. https://doi.org/10.1109/TAFFC.2020.2970399

[18] Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. 2021. pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks. arXiv:2106.09462 [cs.CL]

[19] Edward Raff. 2019. A Step Toward Quantifying Independently Reproducible Machine Learning Research. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/c429429bf1f2af051f2021dc92a8ebea-Paper.pdf

[20] 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* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 784–788. https://doi.org/10.1145/3487553.3524658

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

[22] Seethal. 2022. Seethal/sentiment-analysis-generic-dataset · Hugging Face — huggingface.co. https://huggingface.co/Seethal/sentiment_analysis_generic_dataset. [Accessed 10-02-2024].

[23] Koustuv Sinha. 2023. ML Reproducibility Challenge 2023 | MLRC2023 — reproml.org. https://reproml.org/. [Accessed 03-02-2024].

[24] Sonita Te, Amira Barhoumi, Martin Lentschat, Frédérique Bordignon, Cyril Labbé, and François Portet. 2022. Citation Context Classification: Critical vs Non-critical. In *Proceedings of the Third Workshop on Scholarly Document Processing*, Arman Cohan, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Drahomira Herrmannova, Petr Knoth, Kyle Lo, Philipp Mayr, Michal Shmueli-Scheuer, Anita de Waard, and Lucy Lu Wang (Eds.). Association for Computational Linguistics, Gyeongju, Republic of Korea, 49–53. https://aclanthology.org/2022.sdp-1.6

[25] BI team. 2022. sbcBI/sentiment-analysis-model · Hugging Face — huggingface.co. https://huggingface.co/sbcBI/sentiment_analysis_model. [Accessed 10-02-2024].

[26] Hugging Face Team. 2024. Models - Hugging Face — huggingface.co. https://huggingface.co/models. [Accessed 05-01-2024].

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

[28] Y. Yang, W. Youyou, and B. Uzzi. 2020. Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proceedings of the National Academy of Sciences* 117 (2020), 10762–10768. Issue 20. https://doi.org/10.1073/pnas.1909046117

[29] Abdallah Yousif, Zhendong Niu, James Chambua, and Zahid Younas Khan. 2019. Multi-task learning model based on recurrent convolutional neural networks for citation sentiment and purpose classification. *Neurocomputing* 335 (2019), 195–205. https://api.semanticscholar.org/CorpusID:67867490