Reproducibility Signals in Science: A preliminary analysis

Akhil Pandey Akella
Dept. of Computer Science
Northern Illinois University
aakella@niu.edu

Hamed Alhoori
Dept. of Computer Science
Northern Illinois University
alhoori@niu.edu

David Koop
Dept. of Computer Science
Northern Illinois University
dakoop@niu.edu

Abstract

Reproducibility is an important feature of science; experiments are retested, and analyses are repeated. Trust in the findings increases when consistent results are achieved. Despite the importance of reproducibility, significant work is often involved in these efforts, and some published findings may not be reproducible due to oversights or errors. In this paper, we examine a myriad of features in scholarly articles published in computer science conferences and journals and test how they correlate with reproducibility. We collected data from three different sources that labeled publications as either reproducible or irreproducible and employed statistical significance tests to identify features of those publications that hold clues about reproducibility. We found the readability of the scholarly article and accessibility of the software artifacts through hyperlinks to be strong signals noticeable amongst reproducible scholarly articles.

1 Introduction

Transparency in the scientific process accelerates scientific discovery and strengthens public opinions on scientifically driven matters. Reproducibility plays a crucial role in aiding this transparency, and it is encouraging to have a consensus in the scientific community to address the problem of reproducibility in science. Policymakers, government entities, open source communities, peer-reviewed journals, conferences, and the academic community at large have a shared responsibility to promote reproducible research. Effective dissemination of science cannot happen without trust and integrity in the scientific process. Practically, reproducible science has a first-hand impact in notable places such as research labs, classrooms, industries, and academia. Lack of reproducible research could restrict attaining a deeper understanding of the original researcher's thought process and, therefore, severely impact people involved in the communities mentioned earlier.

The concept of reproducibility is intricate and stratified with different but complementary issues. Before we attempt to understand how to approach the problem of reproducibility, we must first provide some definition of what we mean by this term in this context. Studies such as (Gundersen and Kjensmo, 2018; Cohen et al., 2018; Barba, 2018) highlight how the definition of *reproducibility* varies across different studies and disciplines and how differing definitions can result in confusion. For that reason, the flexible definition presented in Gundersen and Kjensmo (2018) is appealing: "the ability of an independent research team to produce the same results using the same method based on the documentation made by the original research team." Collective efforts from various players of the research community such as publishers, conference organizers, and journals in promoting good practices for ensuring reproducibility in the experimentation process is refreshing, but there is still a lack of agreement on what exactly constitutes a "good practice" which is a concern.

In this study, we attempt to understand the relationship between the structure of science (Thelwall, 2019) and the concept of reproducibility by using statistical significance tests. In doing so, our emphasis is to examine epistemic opacity (Newman, 2015) of linguistic features and structural features concerning reproducibility. We achieve this by running numerous hypothesis tests and identifying the significant factors affecting the reproducibility of scholarly articles. Our goal is to utilize statistical tests to pick signals that could help identify articles requiring more (or less) effort to reproduce.

2 Related Work

Reproducibility is an important concept that affects large communities in general (Mede et al., 2020; Hutson, 2018). The breadth of literature on re-

Proceedings of the 1st Workshop on Information Extraction from Scientific Publications, pages 140–144

producibility spanning different disciplines (Open Science Collaboration, 2012; Prinz et al., 2011; Begley and Ellis, 2012; Peers et al., 2012) has broadly focused on either performing large meta-analyses that reproduce a large set of scholarly articles or qualitative studies that encourage researchers to adopt a certain methodology.

Our study falls in line with the studies that attempts to quantify the factors important for reproducibility, e.g. (Raff, 2019). Identifying such important factors would also be helpful in building machine learning models that can estimate the degree of reproducibility in scholarly articles(Yang et al., 2020).

3 Data

While scientific publications often follow similar structures, there is significant freedom in how ideas are communicated and expressed. This lack of rigidity allows authors to weave stories around fundamental ideas, and the absorption of particular ideas can sometimes be related to how they are presented. We are interested in whether the structure of a publication reveals anything about its potential for (ir-)reproducibility. To examine this, we compiled a collection of scholarly articles that have been evaluated as either reproducible or irreproducible from three different sources. For each article, we gathered comprehensive metadata and extracted structural and linguistic features. These collections of articles include:

- **Brown University**: Collberg et al. (Collberg et al., 2015) conducted a meta-analysis that involved steps in reproducing scholarly articles published in ACM computer science conferences and journals. They found that nearly 50 percent of the examined scholarly articles required extra effort to reproduce the articles. Computer scientists at Brown University led an effort named "Examining Reproducibility in Computer Science" to crowdsource a reexamination of this study (Krishnamurthi, 2015). They performed a meta-analysis of the original study and offered new insights. The data collected provides significant detail about the effort involved in reproducing the studies in the original publications. The current repository provides results for 207 papers; 142 are classified as reproducible and 65 as non-reproducible.

- **Retraction Watch Database (RetractionDB)**: The Retraction Watch Database stores information about scholarly articles that are retracted from conferences and journals (Oransky and Marcus, 2010). It also logs information about the subject/area to which the scholarly article belongs, the country where the article is published, the name of the publisher, the journal name, and most importantly, the reason why the article was retracted. We used this database to find all the scholarly articles in the field of computer science that were retracted under reasons surrounding results not being reproducible, and 34 papers fit these criteria.

- **Badged ACM Papers**: The Association for Computing Machinery (ACM) has introduced badges as a way to signal when publications have been successfully reproduced. We began with 176 articles that were badged as having results reproduced. Of these, 90 were badged as having Reusable Artifacts, and 70 of those had a Functional Artifact badge. We were able to obtain 64 of the papers that had "Results Reproduced" badges and received both a Reusable Artifact and a Functional Artifact badge.

From each of the three sources, we used the available metadata to locate each article. In some cases, we searched by article and authors' names to obtain a DOI or, in some cases, a URL for an article. If we were unable to unambiguously determine this information, the article was dropped from the dataset. Using the DOI, we were able to obtain further metadata and the full text of the article, usually in PDF format. After filling out the metadata and obtaining the full text, we had 305 papers in total; 206 were classified as reproducible, and 99 were classified as non-reproducible. Data and code will be made available as supplementary information upon publishing.

4 Methodology

4.1 Feature Engineering

The motivation for considering the below features stems from the shared intuitions highlighted in (Gundersen et al., 2018; Gundersen, 2020; Raff, 2019) along with checklists from popular publishing venues such as NeurIPS, ICML, etc.

Table 1: List of Structural Features and respective Point Biserial Correlations against target variable

Feature	p-value
Presence of Introduction Section	0.0808
Presence of Methodology Section	0.3112
Presence of Results Section	0.7006
Number of Pages	0.1630
Number of Images	0.3571
Number of Tables	0.7187
Number of Algorithms	0.0654
Number of Hyperlinks	0.0028
Number of Equations	0.4212

1. **Structural features:** Quantitative and qualitative information pertaining to the structure of the scholarly article. This includes information about the existence of particular sections as well as counts of the tables, figures, or algorithms in a given scholarly article. We developed python modules to parse the PDF of the scholarly article in order to extract this information. The features along with respective Point Biserial correlations are mentioned in Table 1.

2. **Linguistic features:** Linguistic indicators quantifying different metrics based on the language used in the scholarly article to differentiate the writing styles of various authors. These indicators include Word count, Average word length, Average sentence length, Frequency of words greater than average word length, Syllable count, and Yule's I measure of lexical diversity (Yule, 2014). These features are general to computational linguistics and are easily understandable. Additionally, we considered metrics such as *Complex words*, which refer to the number of polysyllable words in a given text. This feature was extracted using the python *textblob* library. *Mean Readability* was measured by obtaining the mean of readability metrics such as Flesch Reading Ease Level, SMOG Index, Coleman-Liau index, Automated Readability Index, Dale-Chall Readability Score, Linsear Write Formula, and Gunning FOG. We obtained the values from $textstat$, a python package, to obtain the readability metrics. We also collected the *Sentiment* score for the full text of a given scholarly article and attached a sentiment label (positive = 1, negative = 0) for the respective articles. A similar process was used to obtain the sentiment label for the title of the article.

Table 2: List of Linguistic Features and respective Point Biserial Correlations against target variable

Feature	p-value
Word count	0.5357
Average word length	0.2379
Frequency of words greater than average word length	0.9804
Complex words	0.8394
Syllable count	0.7467
Yule's I measure of lexical diversity	0.1102
Mean Readability	0.0000
Article's sentiment	0.5659
Title's sentiment	0.7335

We gathered this information by implementing python programs that used the python libraries such as *spaCy* and *NLTK* to build the methods for calculating the metrics. All of these linguistic measures were based on the full text of the scholarly article. The features along with respective Point Biserial correlations, are mentioned in Table. 2.

4.2 Point Biserial Correlation

A preliminary statistical analysis of the dependent and independent variables could be performed using correlations. Since our target is a nominal variable, we could not use *Pearson* correlation or *Spearman* correlation as both of them presume the target variable to be continuous. The *point biserial* (Gupta, 1960) correlation matrix measures the correlation between a dichotomous target variable and continuous variables. The results in Table 1 and Table 2 are values obtained by calculating the point biserial correlation coefficient(s) and the associated p-value(s).

4.3 Significance tests

The features mentioned in Tables 1 and 2 are a combination of ordinal and nominal attributes. In order to determine the significance of the features, we had to employ different statistical significance tests such as the *Mann-Whitney U* test (Mann and Whitney, 1947) and *Chi-squared* test (Yates, 1934).

5 Results

We computed correlations and performed statistical significance tests on the combined data sources to identify features that played a significant role in indicating the reproducibility of scholarly articles. The point biserial correlations as shown in Tables 1 and 2 suggested that only **mean readability** and **number of hyperlinks** significantly correlate with reproducibility.

The results of the *Mann-Whitney U* and *Chi-squared* tests show that **mean readability, number of hyperlinks, number of algorithms, average word length, and yule's measure of lexical diversity** to be statistically significant features that align and signal scholarly work that is reproducible with reasonable certainty. More significantly, the readability of a scholarly article and accessibility of software artifacts, either as code repositories, psuedo-code, or algorithms, could be considered strong indicators for reproducibility. It is important to note that these signals do not quantify or assure the reproducibility of a scholarly article but rather help identify articles that require more (or less) effort to reproduce.

Table 3: Mann-Whitney U Significance test for the numerical features

Feature	p-value
Yule's I measure of lexical diversity	0.0131
Word count	0.6547
Average word length	0.0003
Frequency of words greater than average word length	0.9171
Syllable count	0.3910
Complex words	0.9596
Mean Readability	0.0001
Number of Images	0.2039
Number of Tables	0.9586
Number of Algorithms	0.0283
Length of the paper	0.5039
Number of Hyperlinks	0.0011
Number of Equations	0.2148

Our findings were backed by results from statistical experiments such as Point Biserial Correlations, Chi-squared test, and Mann-Whitney U test, and p-values ($p < 0.05$) served as the basis for the significance of our findings. You can obtain a copy of the datasets, experiment setup, and additional software artifacts from Github repository. [1].

Table 4: Chi-squared Significance test for the categorical features

Feature	p-value
Presence of Introduction Section	0.1070
Presence of Methodology Section	0.3728
Presence of Results Section	0.8617
Article Sentiment	0.6646
Title Sentiment	0.8495

6 Discussion

The structure of science involves a well-formed process that begins with factual and valid data, continues through detailed descriptions of experimental procedures, and follows on to clearly presented results. The scientific process has many tenets, but these represent some. They have been promulgated over the years to allow the scientific process to flourish with checks and balances in the form of peer reviews. Contextually, factors such as discipline, year, type of scientific study, etc., play a major role in identifying the effort required to reproduce articles. Therefore, the dataset we built is an essential factor to consider while interpreting our findings that the readability of the scholarly article and accessibility of the software artifacts through hyperlinks are significant features among reproducible scholarly articles. Our motivation is to discover additional latent variables that consider these contextual factors while identifying the effort required to reproduce articles.

7 Conclusions and Future Work

In this study, our pursuit of identifying features that can signal reproducible science involved correlations and significance tests. We found the readability of the scholarly article and accessibility of the software artifacts through hyperlinks to be significant features among reproducible scholarly articles. Our code repository with data and experiments will be available post-publishing.

In the future, we plan on expanding the scope of our study by 1) Gathering more Badged data from ACM; 2) Testing the validity of our findings against adversarial examples; and 3) Observing the effects

[1] https://github.com/reproducibilityproject/reproducibilitysignals

of citing a reproducible article vs non-reproducible ones.

8 Acknowledgement

This work is supported in part by NSF Grant No. 2022443.

References

Lorena A. Barba. 2018. Terminologies for reproducible research.

CG. Begley and LM. Ellis. 2012. *Drug development: Raise standards for preclinical cancer research*. Nature.

K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E. Hunter. 2018. Three dimensions of reproducibility in natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Christian Collberg, Todd Proebsting, and Alex M Warren. 2015. Repeatability and benefaction in computer systems research. *University of Arizona TR*, 14:4.

O. E. Gundersen. 2020. *The Fundamental Principles of Reproducibility*. ArXiv e-prints cs-LG.

O. E. Gundersen, Y. Gil, and D. W. Aha. 2018. On reproducible ai: Towards reproducible research, open science, and digital scholarship in ai publications. *AIMag*, 39(3):56–68.

Odd Erik Gundersen and Sigbjørn Kjensmo. 2018. State of the art: Reproducibility in artificial intelligence. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

S.D. Gupta. 1960. Point biserial correlation coefficient and its generalization. *Psychometrika*, 25:393–408.

M. Hutson. 2018. *Artificial intelligence faces reproducibility crisis*. American Association for the Advancement of Science.

Shriram Krishnamurthi. 2015. Examining reproducibility in computer science.

H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

N. G. Mede, M. S. Sch"afer, R. Ziegler, and M. Weißkopf. 2020. The "replication crisis" in the public eye: Germans' awareness and perceptions of the (ir) reproducibility of scientific research. *Public Understanding of Science*, 9636.

Julian Newman. 2015. Epistemic opacity, confirmation holism and technical debt: Computer simulation in the light of empirical software engineering. In *HaPoC*, volume 487, pages 256–272.

Open Science Collaboration. 2012. An open, Large-Scale, collaborative effort to estimate the reproducibility of psychological science. *Perspect. Psychol. Sci.*, 7(6):657–660.

I. Oransky and A. Marcus. 2010. The retraction watch database.

I. S. Peers, P. R. Ceuppens, and C. Harbron. 2012. In search of preclinical robustness. *Nature reviews Drug discovery*, 10:733.

F. Prinz, T. Schlange, and K. Asadullah. 2011. Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery*, 10.

E. Raff. 2019. A step toward quantifying independently reproducible machine learning research. *Advances in Neural Information Processing Systems 32*, pages 5485–5495.

Mike Thelwall. 2019. The rhetorical structure of science? a multidisciplinary analysis of article headings. *Journal of Informetrics*, 13(2):555–563.

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*, 117:10762–10768.

F. Yates. 1934. Contingency tables involving small numbers and the x2 test. *Journal of the Royal Statistical Society*, 1(2):217–235.

C Udny Yule. 2014. *The statistical study of literary vocabulary*. Cambridge University Press.