PREDICTING THE REPRODUCIBILITY OF SOCIAL AND BEHAVIORAL SCIENCE PAPERS USING SUPERVISED LEARNING MODELS

Jian Wu
Computer Science
Old Dominion University
Norfolk, VA, USA
jwu@cs.odu.edu

Rajal Nivargi
Information Sciences and Technology
Pennsylvania State University
University Park, PA, USA
rfn5089@psu.edu

Sree Sai Teja Lanka
Information Sciences and Technology
Pennsylvania State University
University Park, PA, USA
szl577@psu.edu

Arjun Manoj Menon
Information Sciences and Technology
Pennsylvania State University
University Park, PA, USA
arjunmenon@psu.edu

Sai Ajay Modukuri
Information Sciences and Technology
Pennsylvania State University
University Park, PA, USA
svm6277@psu.edu

Nishanth Nakshatri
Information Sciences and Technology
Pennsylvania State University
University Park, PA, USA
nzn5185@psu.edu

Xin Wei
Computer Science
Old Dominion University
Norfolk, VA, USA
nzn5185@psu.edu

Zhuoer Wang
Computer Science and Engineering
Texas A&M University
College Station, TX, USA
wang@tamu.edu

James Caverlee
Computer Science and Engineering
Texas A&M University
College Station, TX, USA
caverlee@tamu.edu

Sarah M. Rajtmajer
Information Sciences and Technology
Pennsylvania State University
University Park, PA, USA
smr48@psu.edu

C. Lee Giles
Information Sciences and Technology
Pennsylvania State University
University Park, PA, USA
giles@ist.psu.edu

October 22, 2021

ABSTRACT

In recent years, significant effort has been invested verifying the reproducibility and robustness of research claims in social and behavioral sciences (SBS), much of which has involved resource-intensive replication projects. In this paper, we investigate prediction of the reproducibility of SBS papers using machine learning methods based on a set of features. We propose a framework that extracts five types of features from scholarly work that can be used to support assessments of reproducibility of published research claims. Bibliometric features, venue features, and author features are collected from public APIs or extracted using open source machine learning libraries with customized parsers. Statistical features, such as p-values, are extracted by recognizing patterns

in the body text. Semantic features, such as funding information, are obtained from public APIs or are extracted using natural language processing models. We analyze pairwise correlations between individual features and their importance for predicting a set of human-assessed ground truth labels. In doing so, we identify a subset of 9 top features that play relatively more important roles in predicting the reproducibility of SBS papers in our corpus. Results are verified by comparing performances of 10 supervised predictive classifiers trained on different sets of features.

Keywords reproducibility · machine learning · feature extraction · feature selection

1 Introduction

Reproducibility is a defining principle of empirical science. Trust in scientific claims should be based on the robustness of the evidence (e.g., experiments) supporting these claims and subject to verification. Concerns about the reproducibility of published research have gained widespread attention over the past decade, with particular focus on the social and behavioral sciences (SBS), e.g., [1, 2, 3, 4, 5, 6, 7]. By SBS here, we include major disciplines chiefly focused on understanding human behaviors and social systems, such as sociology, political science, economics, and psychology, as well as their respective sub-disciplines, such as computational social science, behavioral economics, social psychology, etc. Research that appears confirmatory when it is in fact exploratory, or makes assertions or predictions that are unlikely given known practical and theoretical limitations, can lead the scientific community and its stakeholders to have inappropriate confidence in reported findings. Manually verifying the reproducibility of scientific claims is non-trivial, usually involving collaboration with original authors and revisiting original experiments, e.g., [4]. These concerns have resulted in a call to arms for greater transparency and openness throughout the research process [8, 9, 10], increased attention to statistical rigor [11, 12], and even the development of automated approaches to assess confidence in existing claims.[1] The work we describe here aims to support the development of computational tools to assist human understanding of the reproducibility, replicability and generalizability of published claims in the SBS literatures.

Specifically, a scientific claim appearing in a research article is a testable, falsifiable assertion typically drawing on existing theories or experiments. In this paper, we focus on claims that represent the authors' final conclusions, drawn based on all studies (theoretical and experimental) in the paper. Below is an exemplar claim from the abstract of a paper recently shared in PsyArXiv [13]. It represents the authors' final conclusions, drawn based on all studies (theoretical and experimental) in the paper.

```
Our results indicate that individuals less willing to engage effortful,
deliberative, and reflective cognitive processes were more likely to believe the
pandemic was a hoax, and less likely to have recently engaged in social-distancing
and hand-washing.
```

Traditionally, the task described above can be treated as a classification problem and can be approached using traditional models such as the bag-of-words or sequence tagging models. However, just incorporating semantic information is insufficient for this task due to the limited information existing in the local representation of a claim itself. Rather, indicators of a claim's reproducibility should be extracted beyond the claim to include broader context, including but not limited to the background of the paper's authors, whether and how the claim is supported by preceding work in the literature, and whether and how the claim is accredited by subsequent work. In this work, we focus on the *conclusive claim* of an SBS paper and assume that its reproducibility can be predicted using global features extracted from the paper. The features we extract are not aligned with a specific claim. However, they offer a representation of a paper's overall profile, which can be useful as the big picture before zooming into claim-dependent reproducibility prediction. Of note, in this work, we focus on feature extraction rather than the prediction model, so we do not distinguish between reproducibility and replicability [14]. We assert that these features can be used for prediction modules in either context, so we use claim reproduciblity to broadly describe the ultimate task.

We propose a software framework that extracts 41 features from SBS papers that can be used for reproducibility assessment. Extracted features include five types – bibliometric, venue-related, author-related, statistical, and semantic. Features are extracted using heuristic, machine learning-based methods, and open source software. Individual feature extractors are tested and evaluated based on manually-annotated ground truth when possible. Here, we focus on describing and evaluating the feature extraction framework and briefly describe the initial utility of these features for reproducibility assessment.

---

[1]E.g., DARPA's Systematizing Confidence in Open Research and Evidence (SCORE) program.

2 Related Works

2.1 Information Extraction Systems

GROBID (GeneRation Of BIbliographic Data) is a machine learning library for extracting, parsing and re-structuring raw documents such as PDF into structured XML/TEI encoded documents with a particular focus on technical and scientific publications [15]. GROBID can be used for extracting metadata from headers, citations, and citation context. Developed based on a hierarchical conditional random field model, GROBID has shown superior performance over many metadata and citation extraction software packages [16, 17]. One feature is to segment the full text of an article to section levels, which helps downstream tasks to quickly locate certain section, such as acknowledgements.

CERMINE (Content ExtRactor and MINEr) is a comprehensive open-source system for extracting structured metadata from scientific articles in a born-digital form [18]. CERMINE is based on a modular workflow designed to extract basic structure (e.g., page segmentation), metadata, and bibliographies. Similar to GROBID, CERMINE segments documents into *metadata*, *body*, *references*, and *other*. The implementations were based on heuristic, e.g., metadata extraction, machine learning, e.g., metadata zone classification and reference string extraction, and an external library iText[2]. In [18], the authors compared the performance of various metadata extraction systems, including GROBID, based on three datasets. CERMINE consistently outperforms GROBID in all datasets in terms of title, abstract, year, and references. On certain datasets, CERMINE outperforms GROBID on certain datasets, such as authors and affiliations, keywords.

PDFMEF (PDF Multi-entity Extraction Framework) is a customizable and scalable framework for extracting multiple types of content from scholarly documents [19]. PDFMEF encapsulates various content classifiers and extractors (e.g., PDFBox[3], academic paper classifier [20], pdffigures2 [21]) for multi-type and scalable information extraction tasks. Users can substitute out-of-box extractors with alternatives.

SCIENCEPARSE[4] is a system to extract structured data from raw academic PDFs. The system was able to extract basic bibliographic header fields from a paper, such as title, authors, and abstract. It can extract and parse citation references and their mentions. It can also segment papers into sections, each with heading and body text. A new version[5] works in a completely different way with fewer output fields but higher quality output.

2.2 Public APIs for Scholarly Articles

Elsevier[6] is an information analytics company which provides a wealth of useful information from books and journals. Along with providing a web browser user-friendly experience, Elsevier also offers APIs to search and retrieve data from their products in a machine readable manner. In this pipeline, the Scopus[7] APIs are used to retrieve bibliometric information of the scholarly articles. Scopus is the largest abstract and citation database of peer-reviewed literature. With over 77.8 million records and 25,100 journal titles from more than 5000 international publishers, Scopus provides research metrics in the fields of science, technology, medicine, social science and arts and humanities. The Scopus APIs allows real-time access articles, authors and institutions in their database.

CrossRef[8] is a not-for-profit association offering an array of services to ensure that scholarly research metadata is registered, linked, and distributed. It interlinks millions of items from a variety of content types, such as journals, books, conference proceedings, working papers, and technical reports. The metadata collected from the members of Crossref can be accessed using the Crossref Metadata Retrieval API[9].

Semantic Scholar[10] is an AI-backed search engine for academic publications. It is designed to highlight the most important and influential papers, and to identify the connections between them. It provides a RESTful API for linking or articles and extracting information from the records on demand.
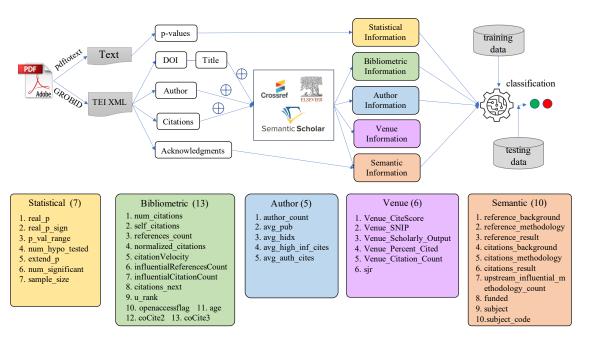
Figure 1: A summary of the feature extraction framework. Different types of features are color-coded.

3 Extraction Framework

3.1 Overview

The top-level architecture of our extraction framework called FEXRep (Feature EXtraction framework for Replicability prediction) is depicted in Figure 1. FEXRep extracts bibliometric, venue, author, statistic, and semantic features from SBS papers. The extraction is performed in two steps. In the first step, we extract raw information directly from content, such as DOI, authors, and citations. In the second step, we derive numerical values to build feature vectors. Our choice of features is based on intuition and promising evidence that certain features may be relevant indicators of reproducibility of findings. We gain deeper insight into their utilities from post-extraction feature analysis.

3.2 Document Preprocessing

The preprocessing step involves parsing unstructured data and generating intermediate metadata that will be used for deriving numerical features. We apply GROBID, which generates a machine-readable XML document under the TEI schema. The machine-readable XML output generated is then further parsed to extract entities, such as address, organizations, authors, and their properties. We use PDFTOTEXT to generate the plain text that is used for extracting p-values.

3.3 Bibliometric Features

This type of features includes quantitative measurements about the target paper's impact in its community. For example, well-researched publications are known to feature a thoroughly curated set of references. Poorly curated bibliographies indicate a lack of cohesiveness in arguments or claims, and insufficient effort invested towards validating ideas against

---

[2]https://itextpdf.com/en

[3]https://pdfbox.apache.org

[4]https://github.com/allenai/science-parse

[5]https://github.com/allenai/spv2

[6]https://www.elsevier.com/

[7]https://www.scopus.com/

[8]https://www.crossref.org/

[9]https://www.crossref.org/services/metadata-retrieval/

[10]https://www.semanticscholar.org/

existing works. Citations made in the context of using or extending methodologies presented in cited work can serve as additional indicators of the robustness of methodology and the reproducibility of claims. Thus, assessing the references and citations of an academic publications is potentially an important source that could help discern reproducible research.

A complete listing of bibliometric features extracted is publically available in an online document[11]. We use the DOI or the title (if DOI is not available) extracted by GROBID as a paper's identifier. Many bibliometric values are obtained by querying digital library APIs, including the Crossref Metadata Retrieval API (hereafter Crossref), Elsevier Scopus API (hereafter Scopus), and Semantic Scholar API[12] (hereafter referred as S2). The records from an API response are refined by calculating string similarties between their titles and the title of the queried paper because in some cases, GROBID returns a partial title. The record whose matching score is greater than 90% is chosen as the final matching result.

num_citations   This metric is the total number of times the target paper is cited since it was published. We use DOIs to query the Scopus API (C_{SC}) and Crossref API (C_{CR}), which return metadata including the citation count and the publication year. The final value is the higher citation count between them. Formally,

$$C(p) = \begin{cases} \max \{C_{SC}(p) \text{ and } C_{CR}(p)\} \\ 0, \qquad\qquad\qquad\qquad\quad \text{otherwise} \end{cases}$$

normalized_citations   This is calculated as the average number of citations per year since the target paper was published. Formally,

$$\overline{C}(p) = C(p)/\Delta Y(p), \quad \Delta Y(p) = Y_{\text{now}}(p) - Y_0(p) \tag{1}$$

in which $Y_{\text{now}}(p)$ and $Y_0(p)$ denote the current year and the publication year of the paper. In rare cases that an API response is not available, a default value of 0 is used.

citation_Velocity   Citation velocity, introduced by S2 in 2016, is an average of the publication's citations for the last 3 years and fewer for publications published in the last year or two, which aims to capture the current popularity and relevance of the work [22]. This metric is pre-calculated by S2 and can be obtained by querying the S2 API using a paper identifier.

citation_next   The time window of 3–5 years after a paper is published is usually considered particularly important for measuring its impact [23]. This feature measures the early citation momentum of a paper. Specifically, this feature is defined as the number of citation a paper receives in the first 3 years after its publication. Formally,

$$\overline{C_3}(p) = \sum_{i=1}^{\Delta Y_3} c_i(p) \bigg/ \Delta Y_3, \Delta Y_3 = \min \{3, \Delta Y(p)\} \tag{2}$$

in which $c_i(p)$ is the number of citation received in year i, obtained by querying the S2 API, and $\Delta Y(p)$ is defined in Eq.(1). The year of publication is obtained from the Crossref API and Scopus API.

influentialCitationCount   Recent work has argued that not all citations are equal, e.g., [24]. In S2, citation metrics are calculated by an algorithm that de-emphasizes absolute citation counts, assigns differential weights to citations depending on citation context, recency, and rate to better determine level of influence. Given a paper identifier, the S2 API returns the number of *influential* citations, which counts citations in which the cited paper had a strong impact on the citing work [25].

references_count   This metric is the number of references the target paper cites obtained from the S2 API and Crossref API, whichever is higher. We consider this feature because it reflects the extend of background and related works the current paper is based on. We set the default value to 0 in case of no API responses.

self_citations   Excessively citing the authors' papers can increase author's h-index, which creates a motivation to strategically use self-citation [26] to promote the apparent impact. Self-citations has been used as a measure to complement h-index [27]. Intuitively, papers that self-cite disproportionately and excessively could potentially reproduce poorly.

Using the extracted author names and references for a given paper, we compute the self-citation count by excluding references authors by any co-author of the target paper. Each author name is parsed to a tuple of (last name, first name

---

[11]`shorturl.at/ghtD2`
[12]https://api.semanticscholar.org/

initial). Two author names match if they have the same first initial and their last names' matching score, calculated by Levenshtein distance, is above a threshold, empirically set to 85%. The self-citation ratio is then calculated as the self-citation count divided by the total number of references. The accuracy of this feature depends on the quality of XML output by GROBID. Errors could be caused by author names that are not extracted from the header or bibliographic sections. By taking the GROBID extraction errors into consideration, the fuzzy matching algorithm results in a root-mean-square-error (RMSE) of 0.09 by comparing automated and manually calculated self-citation ratios for a sample of 37 SBS papers.

openaccessflag Another feature considered is whether the paper has open access. Subscription-based access generally limits the availability of papers. The article being open access can be a potentially important features to observe. This binary feature can be obtained by querying Scopus and Crossref APIs. We assume a paper does not have open access by default.

age This is the number of years since the paper was published

coCite2 The co-citation index between two papers is defined as the number of papers that cite both of them. Papers with higher co-citation indices are usually highly relevant in topics. Therefore, co-citation index can be used for finding topically similar papers. For a target paper p, we use citation graphs to find all "similar" papers with non-zero co-citation indices using the S2 API. This is achieved by first finding all papers (citing papers) that cite the target paper $S_A = \{A_1, \cdots, A_m\}$. Then we find all references in a citing paper A_k: $\{r_1, \cdots, r_l\}$. We next find papers citing r_1: $S_B = \{B_1, \cdots, B_n\}$. The co-citation index between p and r_1 can be calculated as $|S_A \cap S_B|$. This feature counts the numbers of similar papers within 2 years after the target paper was published.

coCite3 This feature is similar to coCite2 except that it counts similar papers within 3 years after the target paper was published.

u_rank Intuitively, the university rank of authors can be used as a indicator of the author's accountability and credibility. We collected university ranking data from the 2020 Times Higher Education rankings[13] and use it as a lookup table to generate the feature value. For a given paper we extract the organization the first and second author (if exists). For matching against the lookup table we use the university the first author is affiliated to. If it is not available we use the second author's affiliation. If neither author's affiliations are available, the default value is used. University names are normalized by removing accents, punctuation marks, and non-ASCII characters. We applied a fuzzy string matcher and set the threshold to 95%, which achieves 100% matching accuracy for 20 random cases with full university names. Another lookup table mapping acronyms to full university names is used in case the latter is not available.

Once matched, a normalized rank between 0 and 1 is calculated as $R_N(u_i) = 1 - R(u_i)/100$, in which $R(u_i)$ is the ranking of university u_i. We consider only the top one hundred universities. If the university's rank is higher than one hundred, we assign $R_N(u_i) = 2$. In cases where there is no match, a default value of 2 is assigned.

3.4 Author Features

Features in this category are related to the authors of the target paper, obtained from S2. Author features include

- *author_count*. The total number of authors of the target paper.
- *avg_pub*. The average number of publications of all authors of the target paper.
- *avg_hidx*. The average h-index of all authors of the target paper.
- *avg_high_inf_cites*. The average number of highly influential citations [25] of all authors.
- *avg_auth_cites*. The average number of citations of all authors.

3.5 Venue Features

Features in this category are pertaining to the conference or journal for a particular paper. All data are obtained from the Scopus API[14] using journal's ISSN as the identifier.

---

[13]https://www.timeshighereducation.com/world-university-rankings
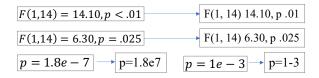[14]https://service.elsevier.com/app/answers/detail/a_id/14834/supporthub/scopus/

Figure 2: Typical cases in which comparison operators are missed when a PDF paper is converted to text.

Venue_CiteScore   CiteScore was first introduced in 2016, as part of an evolving array of research metrics. The metric is a standard to help measure citation impact for journals, book series, conference proceedings and trade journals [28].

Venue_SNIP   The SNIP indicator measures the average citation impact of the publications of a journal, using Scopus data. Unlike the well-known journal impact factor, SNIP corrects for differences in citation practices between scientific fields, thereby allowing for more accurate between-field comparisons of citation impact[15]. SNIP is derived by taking a journal's citation count per paper and dividing it by the citation potential in its subject field [29].

Venue_Scholarly_Output   Scholarly output defines the total count of research outputs, to represent productivity. This feature is calculated as the sum of documents published in a certain venue in the 3 years prior to the current year.

Venue_Percent_Cited   This is calculated as the proportion of documents that have received at least 1 citation.

Venue_Citation_Count   This feature is calculated as the number of citations received in one year for the documents published in the previous 3 years.

SJR   The SJR stands for the SCImago Journal Rank[16], which accounts both the number of citations received by a journal and the importance/prestige of journals where the citations come from. It is calculated as the average number of weighted citations received in a year divided by the number of documents published in the last three years. In case of no API response, a default value of 0 is used.

3.6   Statistical Features

Statistical information is frequently reported in SBS papers when experiments are run. We focus specifically on p-values, a measure of the significance of the observed result. A p-value may serve as an indicator of whether findings from an experiment can be reproduced. In addition, p-values, when presented with the test statistics (e.g., t-test), are especially important references to accept or reject the null hypothesis.

3.6.1   Extracting p-values

To extract p-values, a PDF document is first converted to a text document. We compared several software packages to convert PDF to text, such as XpdfReader, PyPDF2, PDFBox, and PDFMiner and adopted PDFTOTEXT, which produces fewer errors. This is consistent with a recent work on text extractor comparison [30].

Typical errors when extracting p-value expressions include missing comparison symbols such as"=", ">", and "<" (Figure 2), which makes the expression no longer valid. We evaluated the text converter on a random set of 37 papers (hereafter SBS37). PDFTOTEXT successfully converted 90.1% p-values expressions without test statistics (156 out of 173) and 82.5% p-value expressions with test statistics (378 out of 458).

In an SBS paper, p-values can be represented with or without test statistics. The p-values without test statistics are usually in forms of "p <operator> <sign><number>", in which <operator> is one of "=", ">", or "<". The <sign> could be "+" or "−", and the <number> could be an integer (e.g., -2), float (e.g., 0.05), or and exponential (e.g., 1.2e-4). These forms can be captured by regular expressions[17].

The p-values may be reported with test statistics, such as t(12)=4.3, p=0.01 and f(21,30)=2,3, p<0.01, which represent the result of a student's t-test and F-test, respectively. A complete list of p-values patterns in different statistical testings are tabulated in an online document[18]. Using the SBS37 dataset, we compared automatically

---

[15]https://www.journalindicators.com/

[16]https://www.scimagojr.com/SCImagoJournalRank.pdf

[17]Regular expressions are available in the code repository.

[18]shorturl.at/ghtD2

Table 1: Evaluation of p-value and sample size extractors against manually extracted ground truth from PDF and converted text.

DocType	Data Extracted	P	R	F_1
PDF	p-val w/ test stat	0.695	0.920	0.792
	p-val w/o test stat	0.994	0.765	0.864
	Sample size	0.592	0.990	0.741
TXT	w/ test stat	0.698	0.985	0.817
	w/o test stat	0.994	0.926	0.959
	Sample size	0.592	1.000	0.743

extracted p-values against the PDF and converted text. The results (Table 1) indicate that our regular expressions can capture 92% p-values without test statistics from the original PDF, with an overall $F_1 = 0.792$. The precision on capturing p-values with test statistics is 0.994, with an overall $F_1 = 0.864$.

3.6.2 Derived Features From p-values

- *real_p*. A p-value less than 0.05 is usually regarded as a relatively high confidence to exclude the null hypothesis. Because we do not distinguish each hypothesis test, the minimum p-value among all the p-values extracted is used as this feature.

- *real_p_sign*. The signs parsed from p-value expressions. The "<", "=", and ">" are encoded as -1, 0, and 1, respectively.

- *p_val_range*. The p-value range is obtained as the difference of the highest and the lowest p-value in the paper.

- *num_hypo_tested*. We assume the number of hypothesis tests is equal to the total number of p-values with test statistics.

- *extend_p*. A Boolean indicating whether the p-value features are associated with a test.

- *num_significant*. This metric is calculated as the total number of significant p-values (≤ 0.05) including those with and without test statistics as recognized by our parser.

- *sample_size* In an SBS experiment, the sample size is defined as the number of participants or observations. The sample size may explicitly appear in the paper text or can be derived from the p-value test statistic expressions. In one scenario, the sample size could be represented as a integer in free text and is usually noted as $N = N_0$ or $n = n_0$, in which N_0 and n_0 are integers. In another scenario, the sample size can be parsed out by matching the $N = N_0$ pattern in a p-value expression (such as seen in the Chi-squared test). If the $N = N_0$ pattern is missing, the second argument inside χ^2 is treated as the sample size. For certain tests, the sample size can be computed from the test statistic expressions. For example, if a t-test expression is
  `t(df)=number, p<number,`
  then the sample size is `df+1`.

  The sample size extractor is evaluated using the SBS37 corpus, which achieves a high recall (0.990 for PDF and 1.000 for text) but relatively low precision (0.592). The extractor can be improved using the context around an expression to decide whether it includes a sample size.

3.7 Semantic Features

Citation and Reference Intents   A paper could be cited for different reasons. To account for the citation intent, S2 calculates the number of times a given paper is cited as background, methodology, or result [31]. Similarly, citation intent can be obtained for references cited in the given paper. This generates 6 features, namely, *reference_background*, *reference_methodology*, *reference_result*, *citations_background*, *citations_methodology*, and *citations_result*.

upstream_influential_methodology_count   This feature is the number of papers referenced in the target paper in which the citation context is classified as methodology and the referenced paper was classified as influential by S2.

funded   Acknowledgements are ubiquitous in research papers. We consider acknowledgement of a funding agency is a factor for predicting the reproduciblity. We extract acknowledgement organizations using ACKEXTRACT, a framework that distinguishes mentioned and acknowledged entities in a paper [32]. ACKEXTRACT classifies sentences, recognizes all people and organizations from acknowledgement sentences, and then differentiate between acknowledged and

Table 2: Subject distribution of our dataset.

Count	Subject
64	Psychology
30	Sociology and Political Science
22	Linguistics and Language
5	Social Psychology
5	Psychological Science
2	Clinical Psychology
2	Arts and Humanities
2	History and Philosophy of Science
1	Psychiatry and Mental Health
1	Behavioral decision making
1	Philosophy
1	Social Sciences
1	Strategy and Management
1	Developmental Neuroscience
1	Cognitive Neuroscience
139	**Total**

mentioned entities. ACTEXTRACT was evaluated using a corpus of 100 acknowledgement paragraphs containing 146 PEOPLE and 209 ORGANIZATION entities and achieved an overall F_1=0.92.

subject and subject_code In Elsevier, serial titles are classified into 335 *subject fields* by human experts under the All Science Journal Classification (ASJC) scheme. Each subject field is associated with a code ranging from 1000-3700, belonging to 5 *subject areas* – Multidisciplinary, Life Sciences, Social Sciences & Humanities, Physical Sciences, and Health Sciences. We encode the subject field and the subject area returned by the Elsevier Serial Title API into features named *subject* and *subject_code*.

4 Reproducibility Prediction

One important question we attempt to answer is whether it is possible to predict reproducibility from features that can be directly extracted from the paper without redoing the experiments. One existing work used prediction markets to forecast the results of novel experimental designs that test established theories [33]. Another recent study used supervised machine learning models and captured the differences in n-grams between replicating and non-replicating paper. In this study, we treat it as a classification problem and use 10 supervised machine learning models to classify each paper into one of two categories: reproducible vs. non-reproducible [34]. We tried 10 different classifiers to reduce the potential disadvantage of certain classifiers due to the relatively small sample size. Furthermore, we investigate the possibility to reduce the dimensionality of the feature space by removing strongly correlated features.

4.1 Data

We construct a corpus of 139 SBS papers collected from three sources, covering a broad spectrum of subjects (Table 2). These papers have been under careful examination and their reproducibility has been manually labeled by domain experts.

The reproducibility project [2] replicated 99 experimental studies published in three reputable psychology journals (Psychological Science, Journal of Personality and Social Psychology, Journal of Experimental Psychology: Learning, Memory, and Cognition). The results from these replications have been labeled as either replicated or non-replicated and added to our dataset. We also included 12 replication studies from Many Labs 1 [35] and 28 replication studies from Many Labs 2 [36] project. This resulted in a total of 139 labeled papers. A portion of papers used in our work were adopted in a recent reproducibility study [34].

4.2 Feature Extraction

We applied the FEXRep framework to papers in our ground truth dataset and extracted 41 features. As mentioned in Section 3, we apply default values for certain features if the real values are unavailable. To mitigate the artifact
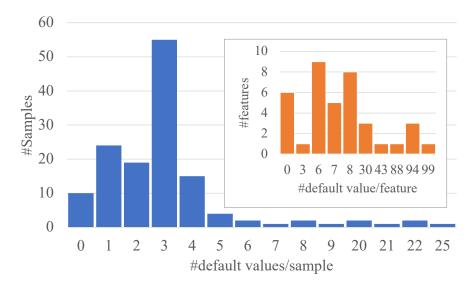
Figure 3: Distribution of the number of default values for each sample (blue), and the number of default values for each feature (orange).

introduced by default feature values, we exclude features if there are less than 15 samples with real values, including real_p_sign (5 samples), p_val_range (14 samples), and sample_size (14 samples). Figure 3 displays the number of samples as a function of default values and the distribution of default values over features. This figure indicates that most samples (96%) have less than 10 features with default values and most features (84%) have less than 30 samples with default values. We refer these 38 features as **core features**.

4.3 Classification Using Core Features

In this section, we apply supervised machine learning models to predict whether a paper is reproducible or not using core features. Our goal is to find out how the prediction accuracy depends on classic machine learning models. The following models were applied.

1. Logistic regression.

2. K-nearest neighbors (k-NN), in which k is set 5.

3. Gaussian process classifier, a non-parametric supervised probabilistic classifier, which assumes that all random variables follow Gaussian distributions. We applied the radial basis function (RBF) as the kernel.

4. Decision tree. Gini impurity is the splitting criterion.

5. Random forest. The max depth is set to 2. The number of estimators is set to 200, and Gini impurity is the splitting criterion.

6. Multilayer perceptron (MLP). MLP is a neural network based supervised classifier that can learn non-linear models.

7. AdaBoost (AB). AB is an ensemble model that fits a sequence of weak learners (i.e., models that are only slightly better than random guessing) on repeatedly modified versions of the data.

8. Naïve Bayes (NB). NB is a probability classifier based on Bayes' theorem with the assumption of conditional independence between every pair of features given the value of the class variable.

9. Quadratic Discriminant Analysis (QDA). QDA uses quadratic surfaces to divide sample points in the feature space.

10. Support vector machine (SVM). The RBF kernel was applied.

Because the sample size is relatively small, we apply a five-fold cross validation (CV) for all models. Figure 4 shows that evaluation results using the core features exhibit significantly different performances. The highest F_1=0.68 is achieved by SVM and QDA, followed by LR with F_1=0.64 and AB with F_1=0.60. SVM and QDA also achieve superior recalls with R=0.99 and R=0.92, respectively. The highest precision is achieved by NB with P=0.64.
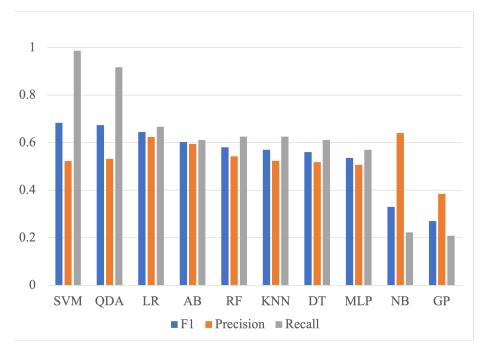
Figure 4: Five-fold CV results using core features sorted by F1 from left to right.

5 Feature Analysis

5.1 Correlations Between Features

The FEXRep framework was designed to extract features that are potentially useful for reproducibility prediction. However, certain features could be correlated with each other, making them less useful in prediction. To capture these correlations, we calculate the Kendall's τ coefficient [37] between any two features with continuous values in our list (features with categorical values are excluded for this analysis). Kendall's τ coefficient ranges from -1 to $+1$. A stronger correlation between two variables results in a higher absolute value. Two random variables without any correlation has $\tau = 0$. We determine feature i and j to be strongly correlated if $\tau_{i,j} > 0.8$. We drop the feature with less real data available. We excluded 6 features that are strongly correlated with at least another feature (Figure 5), including

- *num_significant* (correlated with num_hypo_tested).
- *normalized_citations* (correlated with num_citations)
- *Venue_SNIP* (correlated with Venue_percent_cited)
- *coCite3* (correlated with coCite2)
- *avg_high_inf_cites* (correlated with avg_auth_cite)
- *Venue_CiteScore* (correlated with Venue_percent_cited)

This results in 33 features that are relatively independent with each other. We refer them as **reduced features**. Using *reduced features*, we classified the samples using all machine learning models and obtained consistent results in general (Figure 8).

5.2 Feature Selection

We apply two methods to identify the most relevant features to the reproducibility label: ANOVA-F and Mutual Information (MI).

ANOVA-F ANOVA (analysis of variance) is a parametric statistical hypothesis test for determining whether the means from two or more samples of data come from the same distribution or not. F-test is a class of statistical tests that calculate the variance from two different samples. ANOVA uses F-test to determine whether the variability between
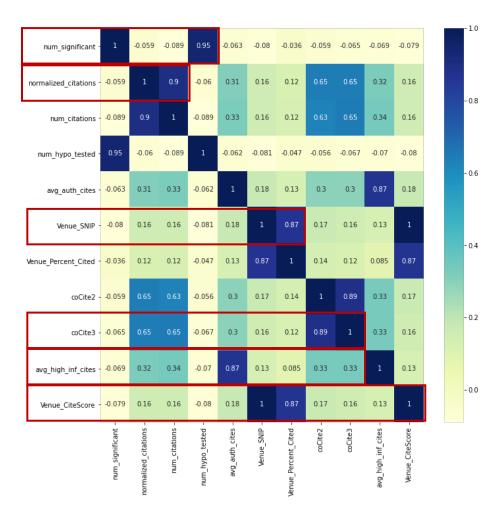
Figure 5: Kendall's τ matrix between highly correlated features. Excluded features are enclosed by red boxes. Note that the matrix is symmetric, so we only study numbers in the lower left triangle.

group means is larger than the variability of the observations within the groups [38]. The larger the value, the more useful the feature is in classification.

Mutual Information (MI)   MI measures the amount of information one can obtain from one random variable given another. The MI between two variables x and y can be calculated as $I(x, y) = H(x) - H(x|y)$, in which $H(x)$ is the entropy for x and $H(x|y)$ is the conditional entropy for x given y. The smaller the value is, the more independent the two variables are. We use non-parametric methods based on entropy estimation from k-nearest neighbors distances to calculate MI [39, 40] . MI values should be non-negative.

5.3   Selecting Top Features

We compared scores calculated using ANOVA-F and MI and show the distributions in Figure 6. We normalized both scores by dividing each value by the maximum value. The shaded region shows top features indicated by ANOVA-F scores (the larger the better). The top features and normalized MI and ANOVA-F scores are tabulated in Table 3.

The MI and ANOVA-F select different sets of relevant features. This is because these two methods captures different types of relations. The ANOVA-F captures linear relationships between variables and the MI captures any types of relationships. As seen in Table 3, there are 3 cross-listed top 10 features identified by both ANOVA-F and MI (in blue text). Among them, the num_hypo_tested feature has MI=0. Estimates of MI can result in negative values due to sampling errors, and potential violation in the assumption that sample rate is high enough for point density to be locally
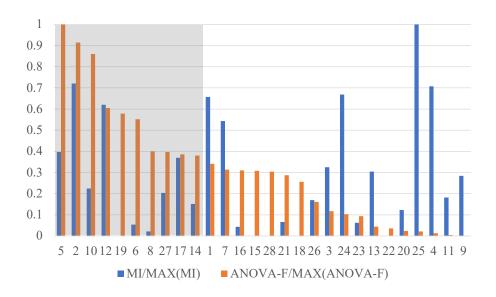
Figure 6: Distribution of mutual information (normalized) and ANOVA-F (normalized) values of *reduced features*. The x-axis labels are core feature IDs used in these calculations.

Table 3: Top 10 features identified by MI (top portion) and ANOVA-F (bottom portion). Feature IDs correspond to x-labels in Figure 6. We show their normalized MI and ANOVA-F values. Blue text are cross-listed features by both MI and ANOVA-F.

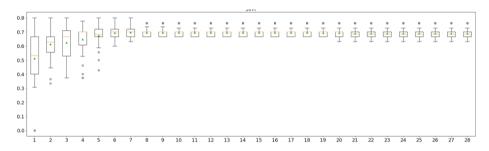| ID	Feature	MI	ANOVA-F
22	Venue_citation_count	0	0.036
18	coCite2	0	0.255
28	age	0	0.305
15	citations_result	0	0.308
19	num_hypo_tested	0	0.578
8	influentialCitationCount	0.021	0.340
16 ●	citations_methodology	0.043	0.310
6	upstream_influential_methodology_count	0.053	0.552
23	Venue_Scholarly_Output	0.062	0.093
21	extend_p	0.066	0.287
5 ●	self_citations	0.397	1.00
2 ●	author_count	0.721	0.915
10 ●	influentialReferencesCount	0.224	0.860
12 ●	reference_result	0.620	0.605
19 ●	num_hypo_tested	0	0.578
6 ●	upstream_influential_methodology_count	0.053	0.552
8 ●	influentialCitationCount	0.021	0.340
27 ●	avg_auth_cites	0.204	0.397
17	citations_next	0.370	0.385
14	citations_background	0.152	0.380

Figure 7: Box-and-whisker plot of SVM F1-measures for each number of selected features identified using ANOVA-F. The green triangles represent arithmetic means. The open dots are outliers beyond caps. The red short lines show medians. The x-labels are feature serial numbers and not feature IDs.
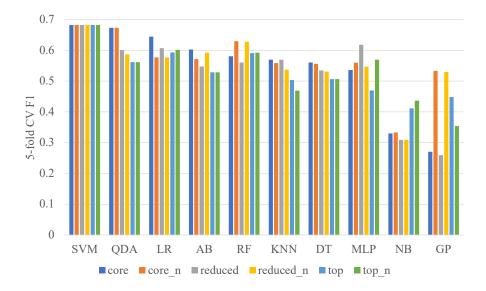

Figure 8: Comparison of F1-measures of classifiers trained on core features, normalized core features, reduced features, normalized reduced features, top 9 features, and normalized top 9 features.

uniform around each point. In our implementation[19], negative MI values are cast to zero. Because of this, we do not select them as top features and focus on features ranked by ANOVA-F.

To investigate how the model performance changes with top features, we evaluate model configurations on classification tasks using repeated stratified 5-fold CV. We choose SVM as the classifier and incrementally add more relevant features selected by ANOVA-F. The box-and-whisker plots are shown in Figure 7. The classifier achieves almost the best performance with the top 8 features identified using ANOVA-F. Adding more features do not seem to help. In fact, adding the last 7 features marginally decreases the performance. We then identify the top 8 features as the most relevant features. To be conservative, we add *citations_methodology* identified using MI as another relevant feature. These 9 features are marked with a orange dot in Table 3.

5.4 Classification with Top Features

To verify top features identified above produce consistent performance across other classifiers, we run the 5-fold CV on all classifiers using the top 9 features selected and compare the performances with classification results using core features and reduced features. We also compare the performance with and without feature normalization. To normalize a feature, we scale a feature to a range between 0 and 1. The transformation is given by $X' = (X - X_{\min})/(X_{\max} - X_{\min})$. The comparison results are illustrated in Figure 8. The F1 of SVM is stable with a marginal decrease with reduced and top features. QDA, LR, AB, KNN, and DT exhibit a general decline when trained with reduced features and top-9

---

[19]`https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html`

features. Normalizing features helps boosting performances in certain conditions. In RF, MLP, NB, and GP, classifiers trained on the top-9 features outperform the core and reduced features. Normalizing feature may or may not boost the performances. Except for QDA, which exhibits a drop of ~ 0.11 and KNN, which exhibits a drop of ~ 0.07, the other classifiers either show a mild decrease (< 0.05) or an increase between core features and top-9 features. The increase of F1 with top-9 features indicate that the classifier may overfit when trained with core features or reduced features, which can be mitigated by adding more training samples. Overall, Figure 8 verifies that the top-9 features selected in Section 5.3 produce generally consistent results across most classifiers except for QDA and KNN. Feature normalization helps to boost performances in certain cases.

6 Conclusion and Discussion

In this work, we develop a framework called FEXRep, which automatically extracts 41 features from SBS research papers. The framework was designed to be modular, scalable, and customizable. New extractors can be added and existing extractors can be updated for better performance. We then use extraction results of this framework to predict the reproducibility of a corpus of 139 SBS papers. By conducting statistical correlation and feature analysis, we finally selected 9 top features, which we believe to be most important. Our work sheds light on the power of using classic machine learning models for evaluating research claims. The normalized top-9 features achieved the best F_1=0.68 using SVM, the best precision of 0.69 using QDA, and the best recall of 0.99 using SVM.

Our study has two limitations. The first is the relatively small sample size. Unfortunately, determining the reproducibility of a claim within a research paper usually requires tremendous effort, rich domain knowledge, and close collaboration, e.g., [4]. With the advocacy and adoption of open science, more papers will be labeled by domain experts, e.g., the repliCATS project [41, 42], and the prediction model will be more robust.

Another limitation is caused by missing values which were set to default values. Lots of default values make us underestimate the true variance of a feature. Most predictive modeling algorithms cannot handle missing data. One simple mitigation is imputing the median for continuous and the modus for discrete predictors. More sophisticated methods to handle missing data build prediction models that estimate missing data.

References

[1] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, 2016.

[2] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251), 2015.

[3] Anna Dreber, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson. Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50):15343–15347, 2015.

[4] Colin F. Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436, 2016.

[5] Richard A Klein, Michelangelo Vianello, Fred Hasselman, Byron G Adams, Reginald B Adams Jr, Sinan Alper, Mark Aveyard, Jordan R Axt, Mayowa T Babalola, Štěpán Bahník, et al. Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490, 2018.

[6] 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. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644, 2018.

[7] Eskil Forsell, Domenico Viganola, Thomas Pfeiffer, Johan Almenberg, Brad Wilson, Yiling Chen, Brian A Nosek, Magnus Johannesson, and Anna Dreber. Predicting replication outcomes in the many labs 2 study. *J. Econ. Psychol.*, 75:102117, 2019.

[8] Brian A Nosek, George Alter, George Christopher Banks, Denny Borsboom, Sara Bowman, Steven Breckler, Stuart Buck, Chris Chambers, Gilbert Chin, Garret Christensen, et al. Transparency and openness promotion (top) guidelines. 2016.

[9] Brian A Nosek, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.

[10] Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie Du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John PA Ioannidis. A manifesto for reproducible science. *Nature human behaviour*, 1(1):1–9, 2017.

[11] Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10, 2018.

[12] Ronald L Wasserstein and Nicole A Lazar. The asa statement on p-values: context, process, and purpose, 2016.

[13] Matthew Stanley, Nathaniel Barr, Kelly Peters, and Ph.D. Seli, Paul. Analytic-thinking predicts hoax beliefs and helping behaviors in response to the covid-19 pandemic. 2020.

[14] Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12–341ps12, 2016.

[15] Patrice Lopez. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'09, pages 473–474, Berlin, Heidelberg, 2009. Springer-Verlag.

[16] Mario Lipinski, Kevin Yao, Corinna Breitinger, Joeran Beel, and Bela Gipp. Evaluation of header metadata extraction approaches and tools for scientific pdf documents. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, pages 385–386, 2013.

[17] Dominika Tkaczyk, Andrew Collins, Paraic Sheridan, and Joeran Beel. Machine learning vs. rules and out-of-the-box vs. retrained: An evaluation of open-source bibliographic reference and citation parsers. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, pages 99–108, 2018.

[18] Dominika Tkaczyk, Paweł Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Łukasz Bolikowski. CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature. *IJDAR*, 18(4):317–335, 2015.

[19] Jian Wu, Jason Killian, Huaiyu Yang, Kyle Williams, Sagnik Ray Choudhury, Suppawong Tuarob, Cornelia Caragea, and C. Lee Giles. Pdfmef: A multi-entity knowledge extraction framework for scholarly documents and semantic search. In *Proceedings of the 8th International Conference on Knowledge Capture*, 2015.

[20] Cornelia Caragea, Jian Wu, Sujatha Das Gollapalli, and C. Lee Giles. Document type classification in online digital libraries. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016*, pages 3997–4002, 2016.

[21] Christopher Clark and Santosh Kumar Divvala. Pdffigures 2.0: Mining figures from research papers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 143–152, 2016.

[22] Keith Kirkpatrick. Search engine's author profiles now driven by influence metrics. *Communications of ACM*, 2016.

[23] Dag Aksnes, Liv Langfeldt, and Paul Wouters. Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open*, 9:215824401982957, 02 2019.

[24] M. Valenzuela, Vu A. Ha, and Oren Etzioni. Identifying meaningful citations. In *AAAI Workshop: Scholarly Big Data*, 2015.

[25] Marco Valenzuela, Vu Ha, and Oren Etzioni. Identifying meaningful citations. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[26] M. Seeber, M. Cattaneo, M. Meoli, and Paolo Malighetti. Self-citations as strategic response to the use of metrics for career decisions. *Research Policy*, 48:478–491, 2017.

[27] A. Kacem, J. W. Flatt, and P. Mayr. Tracking self-citations in academic publishing. *Scientometrics*, 123:1157–1165, 2020.

[28] Jaime A. Teixeira da Silva. Citescore: Advances, evolution, applications, and limitations. *Publishing Research Quarterly*, 36(3):459–468, 2020.

[29] Loet Leydesdorff and Tobias Opthof. Scopus's source normalized impact per paper (snip) versus a journal impact factor based on fractional counting of citations. *J Am Soc Inform Sci Tech*, 61(11):2365–2369, 2010.

[30] Hannah Bast and Claudius Korzen. A benchmark and evaluation for text extraction from PDF. In *2017 ACM/IEEE Joint Conference on Digital Libraries, JCDL 2017, Toronto, ON, Canada, June 19-23, 2017*, pages 99–108. IEEE Computer Society, 2017.

[31] David Jurgens, Srijan Kumar, Raine Hoover, Daniel A. McFarland, and Dan Jurafsky. Measuring the evolution of a scientific field through citation frames. *TACL*, 6:391–406, 2018.

[32] Jian Wu, Pei Wang, Xin Wei, Sarah Rajtmajer, C. Lee Giles, and Christopher Christopher. Acknowledgement Entity Recognition in CORD-19 Papers. In *Proceedings of the 1st Workshop on Scholarly Document Processing*, 2020.

[33] Domenico Viganola, Grant Buckles, Yiling Chen, Pablo Diego-Rosell, Magnus Johannesson, Brian A. Nosek, Thomas Pfeiffer, Adam Siegel, and Anna Dreber. Using prediction markets to predict the outcomes in the defense advanced research projects agency's next-generation social science programme. *Royal Society Open Science*, 8(7):181308, 2021.

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

[35] Richard A Klein, Kate A Ratliff, Michelangelo Vianello, Reginald B Adams Jr., Štěpán Bahník, Michael J Bernstein, Konrad Bocian, Mark J Brandt, Beach Brooks, Claudia Chloe Brumbaugh, and et al. Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, 45(3):142–152, 2014.

[36] Richard A Klein, Michelangelo Vianello, Fred Hasselman, Byron G Adams, Reginald B Adams Jr, Sinan Alper, Mark Aveyard, Jordan R Axt, Mayowa T Babalola, Štěpán Bahník, et al. Many labs 2: Investigating variation in replicability across samples and settings. *AMPPS*, 1(4):443–490, 2018.

[37] M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1-2):81–93, 06 1938.

[38] Max Kuhn and Kjell Johnson. *Feature Engineering and Selection: : A Practical Approach for Predictive Models*. Chapman and Hall/CRC, 2019.

[39] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004.

[40] Brian C. Ross. Mutual information between discrete and continuous data sets. *PLOS ONE*, 9(2):1–5, 02 2014.

[41] Anca Hanea, David P Wilkinson, Marissa McBride, Aidan Lyon, Don van Ravenzwaaij, Felix Singleton Thorn, Charles T Gray, David R Mandel, Aaron Willcox, Elliot Gould, and et al. Mathematically aggregating experts' predictions of possible futures. Feb 2021.

[42] Brian A Nosek, Tom E Hardwicke, Hannah Moshontz, Aurélien Allard, Katherine S Corker, Anna Dreber, Fiona Fidler, Joseph Hilgard, Melissa Kline Struhl, Michele B Nuijten, and et al. Replicability, robustness, and reproducibility in psychological science. Feb 2021.