The Pennsylvania State University

The Graduate School

AN ASSESSMENT OF REPRODUCIBILITY OF SOCIAL AND BEHAVIORAL SCIENCE PAPERS USING SUPERVISED LEARNING MODELS

A Thesis in

Computer Science and Engineering

by

Rajal Nivargi

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science

August 2021

The thesis of Rajal Nivargi was reviewed and approved by the following:

Clyde Lee Giles

David Reese Professor of Information Sciences and Technology

Thesis Advisor

Sarah Rajtmajer

Associate Professor of Information Sciences and Technology

Rui Zhang

Assistant Professor in the Computer Science and Engineering Department

Jian Wu

Assistant Professor of Computer Science Old Dominion University

Special Signatory

Chitaranjan Das

Distinguished Professor of Computer Science and Engineering

Head of the Department of Computer Science and Engineering

Abstract

In the last decade, there has been increased conversation over the "reproducibility crisis" and "replication crisis" in various medical, life and behavioral sciences [1]. This thesis focuses on the social and behavioral sciences(SBS) research claims. We try to assess prediction of reproducibility of SBS papers using supervised machine learning models. We use a framework of feature extraction to retrieve 5 categories of features namely: bibliometric features, venue features, and author features from public APIs or open source machine learning libraries with customized parsers, Statistical features by recognizing patterns in the body text and semantic features from public APIs or using natural language processing models. These features are analysed using different feature selection methods such as pairwise correlations, mutual information and ANOVA-F values. Their importance for predicting a set of human-assessed ground truth labels for the SBS papers was studied. We identify the top features based on the feature selection methods by comparing the performance of 10 supervised machine learning models.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to take this opportunity to thank all the people who have helped me throughout my graduate program and successfully completing my thesis. I am grateful for the opportunity to be a part of the SCORE team. It has been an great learning experience. I would like to thank my advisors, Professor Lee Giles and Professor Sarah Rajtmajer for their support and guidance throughout my time with the team. I am grateful to Professor Jian Wu for his invaluable advice and help for my research. I would like to acknowledge the contributions of my team members: Arjun Manoj Menon, Sree Sai Teja Lanka, Sai Ajay Modukuri, Xin Wei and Zhuoer Wang in the feature extraction pipeline. Special thanks to my committee member Rui Zhang for participating in my defense. I would also like to appreciate the constant inspiration from my parents, friends and family. Without everyone's encouragement and assistance, this wouldn't have been possible.

Finally, I acknowledge that this material is partially supported by Defence Advanced Research Projects Agency cooperative agreement No. W911NF-19-2-0272. The findings and conclusions represented in this work do not necessarily reflect the view or the policy of the Government, and no official endorsement should be inferred.

Chapter 1
Introduction

1.1 Background

Scientific research studies play an important role in our lives by defining policies, navigating our complex world and propel development of mankind. Reproducibility and replication are pillars of trust in scientific research. Leek and Jager [2] defined a study to be replicable "if an identical experiment can be performed like the first study and the statistical results are consistent" and reproducible "if all of the code and data used to generate the numbers and figures in the paper are available and exactly produce the published results". A research study may have irreproducible findings/evidence because of random or systematic error, differences between replication and original data or results from the study being false positive or the replication results being false negative. [3] Recent studies in fields such as medicine [4–6], economics [7–9], neuroscience [10] and psychology [11–14] among others. This has led to the term of reproducibility or replication crisis in scientific research. [1, 2, 15]. In a survey conducted by Nature in 2016, 90% of the respondents were convinced that there is a 'reproducibility crisis' [16].

In this thesis, we focus on the research area of social and behavioral sciences. This area applies to a varied range of disciplines such as anthropology, sociology, and psychology.

It involves careful analysis of human behavior. The claims made by such scientific research as hypotheses are tested using empirical studies. Recently, there has been a gap between the original and replicated studies. [3, 17, 18]. For example, The Reproducbility projects, coordinated by Center for Open Science have been replicating the entire studies including data collection and analysis. Out of the 100 papers, 61 did not replicate. The replicated studies provided weaker evidence for the original results despite using the same methods. This outcome is noticed by other studies as well. 3 out of 13 papers failed the Many Labs replication project. [19] Though the studies chosen were well-known to be robust and of high quality, these have been infrequently replicated or not at all. This makes it important to understand the different factors that contribute to the study being reproducible.

1.2 Motivation

Scientific research papers attempt to transparently describe the methodology and evidence used to bolster their claims. Scientific research is manually peer-reviewed by domain experts participating as review committees at conferences and for journals. This process can differ according to the reviewer expertise, venue requirements and research domains. The process does not take reproducibility of the paper into account because of the limited time and accessibility of resources available at the time of review. The reviewers have to rely on the theoretical or methodological reasons presented in the paper. Though the research may seem confirmatory from their explanations, it may not necessarily be reproducible. This instills false confidence in inaccurate research because of practical or theoretical limitations.

Manually verifying claims for reproducibility is non-trivial and resource intensive. It requires access to the data or data collection pipelines to acquire the same input data. The experiments to be run on this data can be expensive, time-consuming or

hard-to-access. [20] The methods for analysis and testing used in the paper may not be completely covered in the paper. Research experiments may have underlying details and steps not outlined in the paper. This will need the involvement of the original authors. This back-and-forth between the authors and reviewers can also result in some delays in replicating the work. This has emphasized the need of increased transparency throughout the research process to asses the confidence in existing claims. In this thesis, we have made an attempt to automate the process of validating the credibility of research claims. This aims to ensure higher speeds and lower costs than manual evaluation.

1.3 Approach

In this work, we target the prediction of reproducibility of social and behavioral sciences(SBS) research articles. Supervised machine learning models are observed to have potential to aid the process of reviewing the scientific articles over human effort. We performed experiments on a dataset of 139 papers which are labelled as reprodicible or not. A total of 41 features were extracted using a modular, scalable and customizable feature extraction framework called FEXRep. We compared the performance of 10 supervised machine learning models on different sets of features in the dataset. These sets were chosen by feature selection based on Correlation of features using Kendall Tau coefficient, Mutual information and ANOVA-F values. A set of 9 features was chosen to be the top features. Finally, the model's interpretability was studied by using a model-agnostic technique called LIME [21].

Chapter 2
Related work

The credibility of scientific research depends on its reliability and efficiency. Recently, there have been efforts to develop tools to assist human understanding of the reproducibility, replicability and generalizability of literature. This has given birth to a field of metascience - 'the scientific study of science itself' [22]. Referring to the same kind of work, 'metaresearch' is defined as 'an evolving scientific discipline that aims to evaluate and improve research practices. It includes thematic areas of methods, reporting, reproducibility, evaluation, and incentives (how to do, report, verify, correct, and reward science).' [23]. As the number of scientific publications grow, the need and opportunity to study the research practices and evidences in these publications grow as well. In 2015, over 850 meta-science publications were identified in a short period of January to May [23]. The most notable reproducibililty projects and replications studies are:

1. The Reproducibility Project: Psychology coordinated by Center of Open Science [3] This project replicated 100 papers from three psychology journals using the original materials(when available) in the year 2008. Only 39% of the total papers were said to have replicated the original work. Though the study does not state a single indicator of replication success, it did offer the conclusion that a large number of papers offer weak evidence to the original findings.

2. The Many Labs Project [17] This project performed a large scale replication of 13

classic and contemporary psychological effects such as anchoring, sunk cost bias and priming, among others, with 36 samples and settings. The project demonstrated that 11 out of 13 effects can be replicated(in terms of statistical significance). It investigated the dependence of the effect on different methods of data collection.

3. The Many labs 2 Project [18] This project conducted replications of 28 classic and contemporary findings out of which 15 were said to be replicated based on evidence of statistically significant effect. They examined the variation in effect magnitudes across samples and settings.

4. The Social Science Replication Project(SSRP) [19] This project studied the replication of 21 experimental studies in social sciences published in Nature and Science between 2010 and 2015. Out of these, 13 papers were shown to be replicable. The study concluded that the predictions of replication are not a result of chance alone and efforts can improve over time.

Finding the factors that help explain if the paper is reproducible or not, is important. Studies [24] have noted that the median power (statistical significance) of published studies is often below 50%, assuming that all effects under study are true and accurately estimated. Besides statistical significance, the citations of individual studies can also supporting evidence for the study being replicable [25].

Three approaches to predict replication outcomes are: surveys, structured protocols presented to small groups and prediction markets with participants betting on the replication success by buying and selling contracts for replication. The prediction markets have been shown to obtain accurate forecasts of the outcome of replicability. They were able to predict 29/41 (71%) of the replications in [26] However, the individuals involved would have to be experts in the field of given research. It is helpful to anticipate the likelihood of a paper being replicable or not in advance of human effort.

There have been some experiments which use machine learning to predict the repro-

ducibility of research articles [25, 27]:

1. Black box statistical models to predict replicability [28]: In this paper, the authors attempted to predict the replicability of research from features including statistical features(P-value, effect sizes), author features(number of authors, their citations), citations of research, funding, etc. They approximated the accuracy of 70% using a Random Forests model for a small labelled dataset of 131 direct replications. Though this was a significant result as compared to the prediction markets [26], the feature extraction was a manual process which can be a non-trivial process for an individual with limited domain understanding.

2. Estimating deep replicability using human and artificial intelligence [29]: This work uses the word2vec [30], neural network based method with standard settings to collect the features from the paper by defining the qualitative relationship(co-occurrence) of each word in the corpus. This paper-level vector representing the unique linguistic information was using to train a simple ensemble model of bagging with random forests and bagging with logistic. They were able to get the accuracy of model predictions between 0.65 to 0.78.

3. Probabilistic forecasting [31] This paper uses a different method of probabilistic forecasting using the original studies information and the replication studies sample size only. They were able to predict the effect estimate of the replication study for two out of four data sets from different areas of research.

In this thesis, we will be attempting to further the attempts to predict the reprodicibility of similar datasets tested by above articles using automatically extracted features to train the machine learning models.

Chapter 3
Feature Extraction Pipeline

The Feature EXtraction framework for Replicability prediction(FEXRep) is used to extract the features for the dataset. [32]. This framework extracts 5 categories of features: bibliometric, venue, author, statistic and semantic features from Social and Behavioral Sciences(SBS) papers. The works elaborated in Chapter 2 discuss the importance of the categories of features. The bibilometric, author and venue features were used in training the black box statistical models in [28]. The semantic features are similar to the word embeddings used in [29]. Statistical features are deemed of importance in prediction of replicability [3, 31, 33, 34].

A research paper is fed in the pipeline and pre-processed. The next step is extracting raw information from the text such as Digital Object Identifier(DOI), author names, among others. Different methods are used after this step to get numerical values from publicly available scholarly APIs, paper text and derived information. The overall summary of the feature extraction pipeline is shown in 3.1. This builds a feature vector to be used to train the machine learning models in the experiments.

3.1 Document Pre-processing

GROBID is used at this step of the pipeline to parse the research article. (The research articles are downloaded as PDF files beforehand). GROBID (GeneRation Of BIbliographic

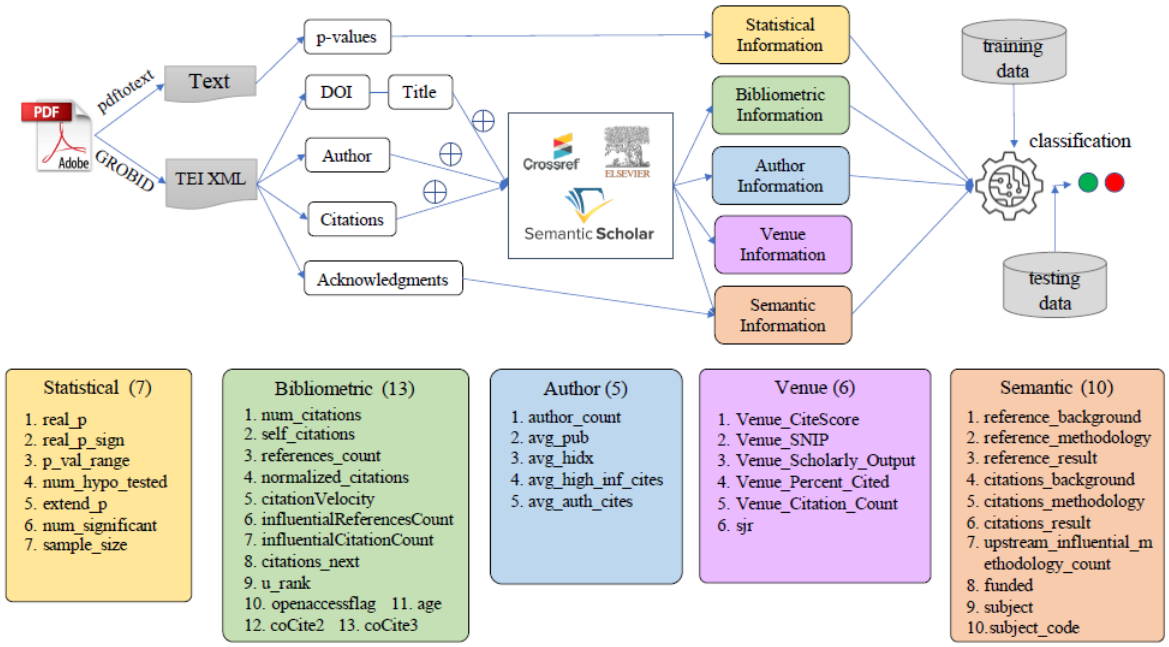


Figure 3.1: Summary of feature extraction pipeline

Data) is a machine learning library. It is used for extracting, parsing and re-structuring raw documents such as PDF into structured XML/TEI encoded documents particularly on technical and scientific publications [35]. It is used to extract metadata from headers, citations, and citation context. It segments the full text of the article into section levels which are used in the next steps of the pipeline. Other tools that were available for the same task include CERMINE (Content ExtRactor and MINEr) [36], PDFMEF (PDF Multi-entity Extraction Framework) [37] and ScienceParse[1]. GROBID has shown best performance in terms of accuracy and F1 score in extracting metadata from articles using out-of-the-box tools [38, 39]. PDFTOTEXT is used to generate the plain text that is used for extracting p-values.

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

3.2 Bibliometric features

Bibliometrics is the use of statistical methods to explore the impact of a research article in that particular field. Citations and references are considered the most important features of bibilometric study. They help correlate different articles that use the same or extended methodologies to serve as indicators of robustness of that method. Thus, assessing citations of the particular article is a way of understanding the influence of the paper in that area of research. It can be a source of identifying reproducible research. The three publicly available scholarly APIs used to query this information are:

1. Elsevier[2] 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[3] 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.

2. CrossRef[4] 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

[2]https://www.elsevier.com/
[3]https://www.scopus.com/
[4]https://www.crossref.org/

API[5].

3. Semantic Scholar[6] 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.

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[7] (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. DOIs is used 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,

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

normalized_citations This metric is calculated using the num_citations feature above and the publication year of the target paper retrieved from the Crossref and Scopus API. Their ratio gives the average number of citations per year since the target paper was published. Formally,

$$\overline{C}(p) = C(p)/\Delta Y(p), \quad \Delta Y(p) = Y_{\text{now}}(p) - Y_0(p) \tag{3.1}$$

[5]`https://www.crossref.org/services/metadata-retrieval/`
[6]https://www.semanticscholar.org/
[7]https://api.semanticscholar.org/

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 [40]. This metric is a feature directly retrieved from the Semantic Scholar(S2) API using a paper identifier(DOI).

citation_next The time window of 3–5 years after a paper is published is usually considered particularly important for measuring its impact [41]. 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,

$$\overline{C_3}(p) = \sum_{i=1}^{\Delta Y_3} c_i(p) \Big/ \Delta Y_3, \Delta Y_3 = \min\{3, \Delta Y(p)\} \tag{3.2}$$

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.(3.1). 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., [42]. 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 [43].

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 [44] to promote the apparent

impact. Self-citations has been used as a measure to complement h-index [45]. 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 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[8] 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 $R_{\mathrm{N}}(u_i) = 1 - R(u_i)/100$, 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 $R_{\mathrm{N}}(u_i) = 2$. In cases where there is no match, a default value of 2 is assigned.

3.3 Author features

The next type of feature set that are related to the bibliometirc features are author features. The research articles written by reputed authors tend to have higher citation counts and thus, higher impact in the research area because of their past success and a trust in the quality of research by the author [46].:

[8]https://www.timeshighereducation.com/world-university-rankings

1. *author_count*. The total number of authors of the target paper. [47]

2. *avg_pub*. The average number of publications of all authors of the target paper.

3. *avg_hidx*. The average h-index of all authors of the target paper.

4. *avg_high_inf_cites*. The average number of highly influential citations [43] of all authors.

5. *avg_auth_cites*. The average number of citations of all authors.

3.4 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[9] using journal's ISSN as the identifier. The venue a certain paper has been published at is important since it demonstrates the paper having undergone structured peer review process by the journal/conference committee of experts. This helps improve trust in the paper and possibly predict its reproducibility. [48]

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 [49].

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[10]. SNIP is derived by taking a journal's citation count per paper and dividing it by the citation potential in its subject field [50].

[9] `https://service.elsevier.com/app/answers/detail/a_id/14834/supporthub/scopus/`
[10] https://www.journalindicators.com/

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[11], 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.

3.5 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. [3] In addition, p-values, when presented with the test statistics, are especially important references to accept or reject the null hypothesis.

3.5.1 Extracting p-values

To extract p-values, a PDF document is first converted to a text document. After comparison of several software packages such as XpdfReader, PyPDF2, PDFBox, and PDFMiner, it was observed that PDFTOTEXT produces fewer errors. This is consistent with a recent work on text extractor comparison [51].

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

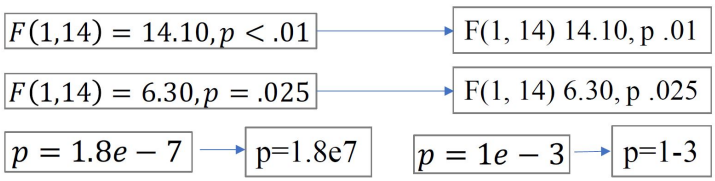


Figure 3.2: Typical cases in which comparison operators are missed when a PDF paper is converted to text.

Typical errors when extracting p-value expressions include missing comparison symbols such as"=", ">", and "<" (Figure 3.2), 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[12].

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[13]. Using the SBS37 dataset, we compared automatically extracted p-values against the PDF and converted text. The results (Table 3.1) 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 $F_1 = 0.864$.

[12]Regular expressions are available in the code repository.
[13]`shorturl.at/ghtD2`

Table 3.1: 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

3.5.2 Derived Features From p-values

1. *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.

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

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

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

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

6. *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.

7. *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.

3.6 Semantic features

The semantic features are indirect features extracted from the text of the paper. These are converted to numerical values such as binary indicators, categorical values and counts.

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 [52]. 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 [53]. AckExtract classifies sentences, recognizes all people and organizations from acknowledgement sentences, and then differentiate between acknowledged and 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*.

Chapter 4
Experiments

The purpose of this thesis is to assess if it is possible to predict the reproducibility of a research article based on the features described in chapter 3.The distribution of labelled data into the two classes is shown in the figure 4.1. This is treated as a binary classification problem in this thesis. We use supervised machine learning models to classify each paper as reproducible or not. Different sets of features were tested to predict the reproducibility. This helps understand which features may be important for this prediction.

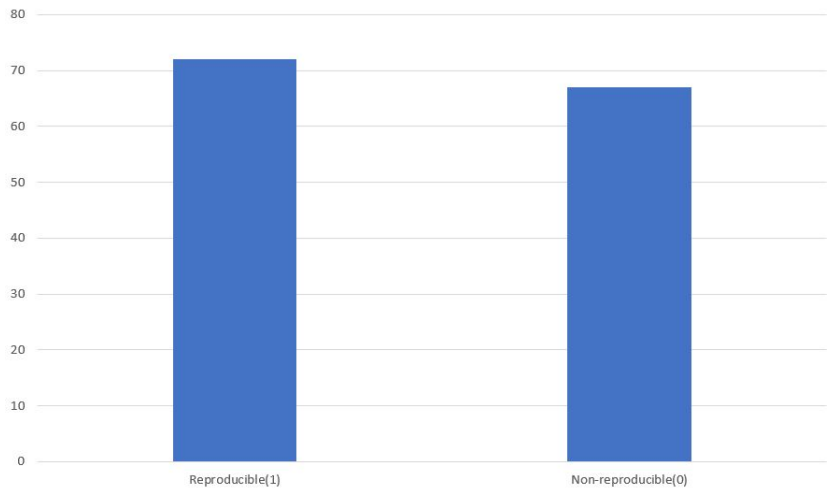


Figure 4.1: Distribution of labels

Table 4.1: 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**

4.1 Data

The dataset used for this thesis is a corpus of 139 research papers in the Social and Behavioral Sciences(SBS) field. This was collected from the three data sets covered in chapter 2. The reproducibility project 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 and 28 replication studies from Many Labs 2 project.These provide papers which are labelled as reproducible or non-reproducible by domain experts. The papers were scraped to collect them as PDF files. The papers cover a broad spectrum of subjects (extracted from the SCOPUS API) as shown in Table 4.1

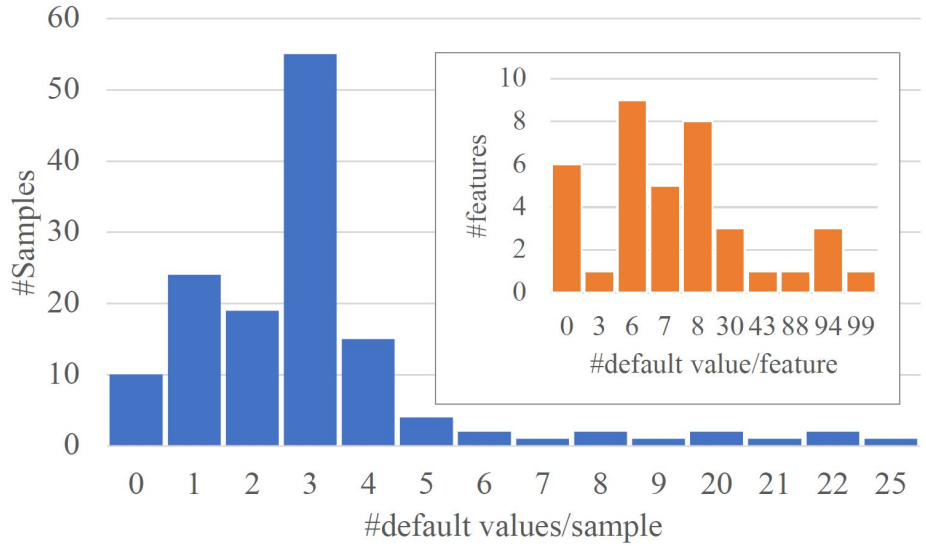


Figure 4.2: Distribution of the number of default values for each sample (blue), and the number of default values for each feature (orange).

4.2 Features

The FEXRep framework elaborated in chapter 3 was used on this dataset. A total of 41 features were extracted for each paper. In some cases, where the framework is unable to retrieve values for the features, a default value is assigned to that feature. This may be an issue while predicting the reproducibility of a paper with many default values since they do not represent the actual features. To mitigate this, we first need to understand the distribution of the default values in the data set extracted from the framework.

Figure 4.2 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**. We exclude 3 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).

4.3 Methods and Evaluation

As discussed above, we apply supervised machine learning models to predict whether a paper is reproducible or not using the features extracted for our dataset. These are used off the shelf from the scikit-learn library. [54].The sample size we use is relatively small, so 5-fold cross validation is used on all models. The following models were applied.

1. Logistic regression with liblinear solver: Fit for smaller datasets and high dimensionalilty as it uses a coordinate descent algorithm.

2. K-nearest neighbors (k-NN), in which k i.e. the number of neighbours 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 with Gini impurity as 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): A neural network based supervised classifier that can learn non-linear models.

7. AdaBoost (AB): 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): 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): It uses quadratic surfaces to divide sample points in the feature space.

10. Support vector machine (SVM) with RBF kernel

The values of different features in the dataset can have different value ranges. Standardization of the data set, in general, benefits the machine learning algorithms. Thus, we preprocess the data by normalization. Normalization is a process of scaling individual samples to have unit norm. For this, We use the function from the scikit learn library [54] on our data.

As we train the classification predictive model, we want to assess how good it is. There are many different ways of evaluating the performance. In this thesis, we use 4 evaluation metrics from the scikit-learn library [54]: Accuracy classification score, F1 score, Precision and Recall. For the cross-validation(CV) methods, a score computed at each CV iteration is the score method of the estimator. A mean and standard deviation of this score for each model is also considered for evaluation.

4.4 Feature Analysis

4.4.1 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 [55] between any two features with continuous values in our list. The features with categorical values are excluded for this analysis. These include 'funded', 'openaccessflag', 'subject' and 'subject_code'. The correlation matrix for features is shown in figure 4.3.

Kendall's τ coefficient ranges from -1 to $+1$. A stronger correlation between two

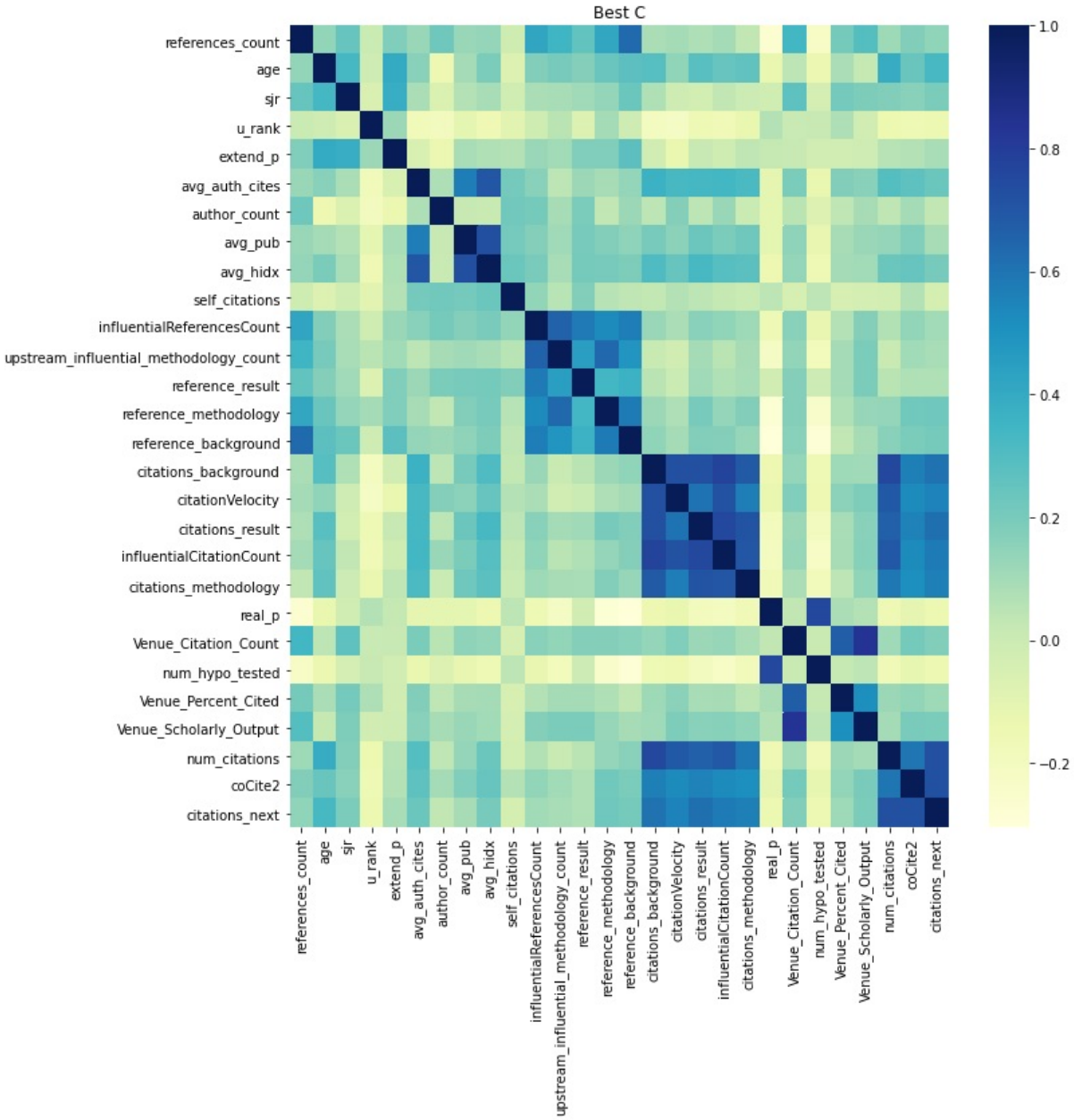


Figure 4.3: Correlation matrix

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 4.4), including

1. *num_significant* (correlated with num_hypo_tested).

2. *normalized_citations* (correlated with num_citations)

3. *Venue_SNIP* (correlated with Venue_percent_cited)

4. *coCite3* (correlated with coCite2)

5. *avg_high_inf_cites* (correlated with avg_auth_cite)

6. *Venue_CiteScore* (correlated with Venue_percent_cited)

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

4.4.2 Mutual Information

Mutual Information(MI) is a statistical metric that measures the amount of information one can obtain from one random variable given another. Mutual information is always larger than or equal to zero, where the larger the value, the greater the relationship between the two variables. If the calculated result is zero, then the variables are independent. The target variable in this case is the label of prediction i.e. 1 for reproducible or 0 for non-reproducible. We are trying to understand the mutual information between each feature values and the paper's label.

4.4.2.1 Method 1

The MI function from the scikit-learn [54] library is used to find the dependence or "mutual dependence" between two random variables. It is calculated by non-parametric methods based on entropy estimation from k-nearest neighbors distances. [56, 57]

4.4.2.2 Method 2

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.

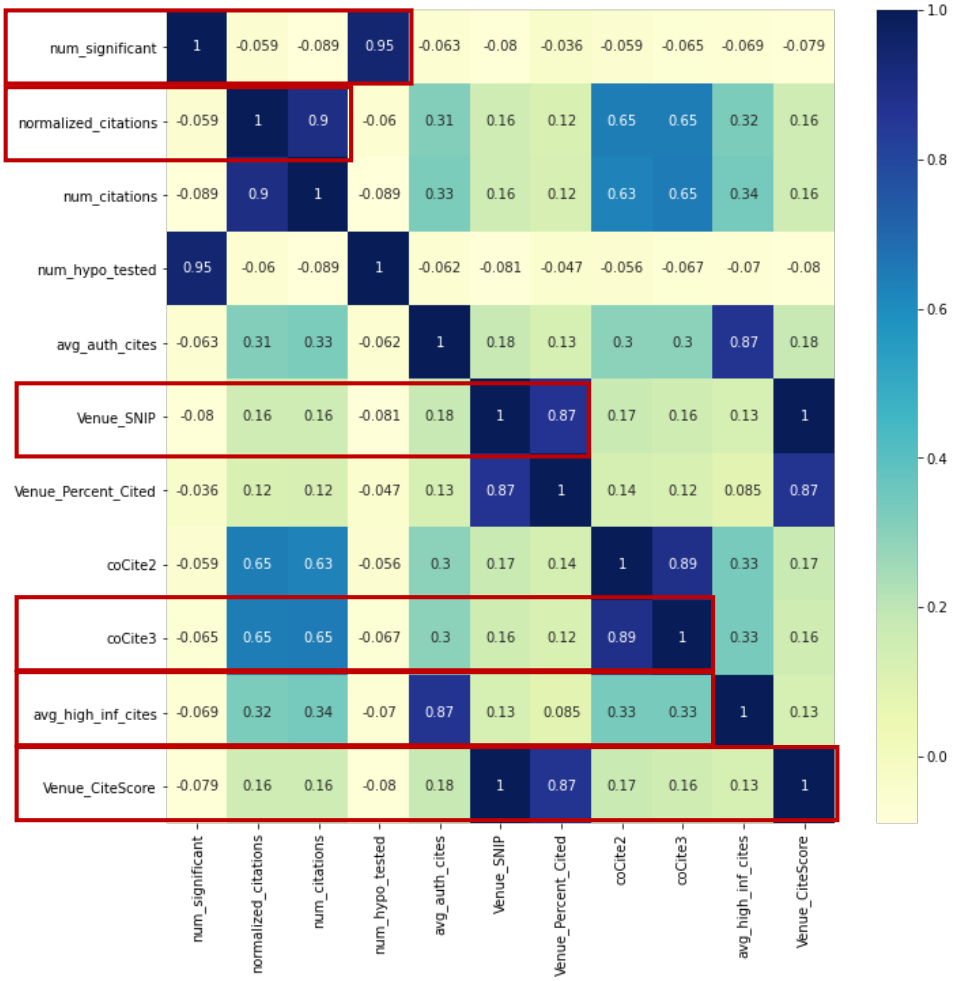


Figure 4.4: Kendall's τ matrix between highly correlated features. Excluded features are enclosed by red boxes. Note that the matrix is symmetric, so we only study numbers in the lower left triangle.

The smaller the value is, the more independent the two variables are. The mutual information function from the pyitlib library[1] is used for calculating the normalised mutual information between arrays X and Y, each containing discrete random variable realisations.

[1]https://pafoster.github.io/pyitlib/

4.4.3 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 group means is larger than the variability of the observations within the groups [58]. The larger the value, the more useful the feature is in classification. In this case, the target variable is the label of reprodicbility of the paper i.e. 1 for reproducible or 0 for non-reproducible.

4.4.4 Feature selection

The independent features were used for feature selection methods described above. The SelectKBest function from scikit-learn library [54] was used to select the top features in each method. This was based on the target variable of 'label' for each sample. This functions removes all features but the k-highest features i.e. features with highest score based on the method of selection. The methods based on F-test estimate the degree of linear dependency between two random variables. On the other hand, mutual information methods can capture any kind of statistical dependency. A comparison between the scores predicted by the three methods are shown in the figure 4.5. The scores from each method for the features has been normalized for the sake of this comparison using MINMAX normalization. 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 type of statistical relationship.

To select the number of features to be chosen for further analysis and prediction, we choose SVM as the classifier and incrementally add more relevant features selected by each method. The box-and-whisker plots are shown in Figure 4.6. The classifier achieves

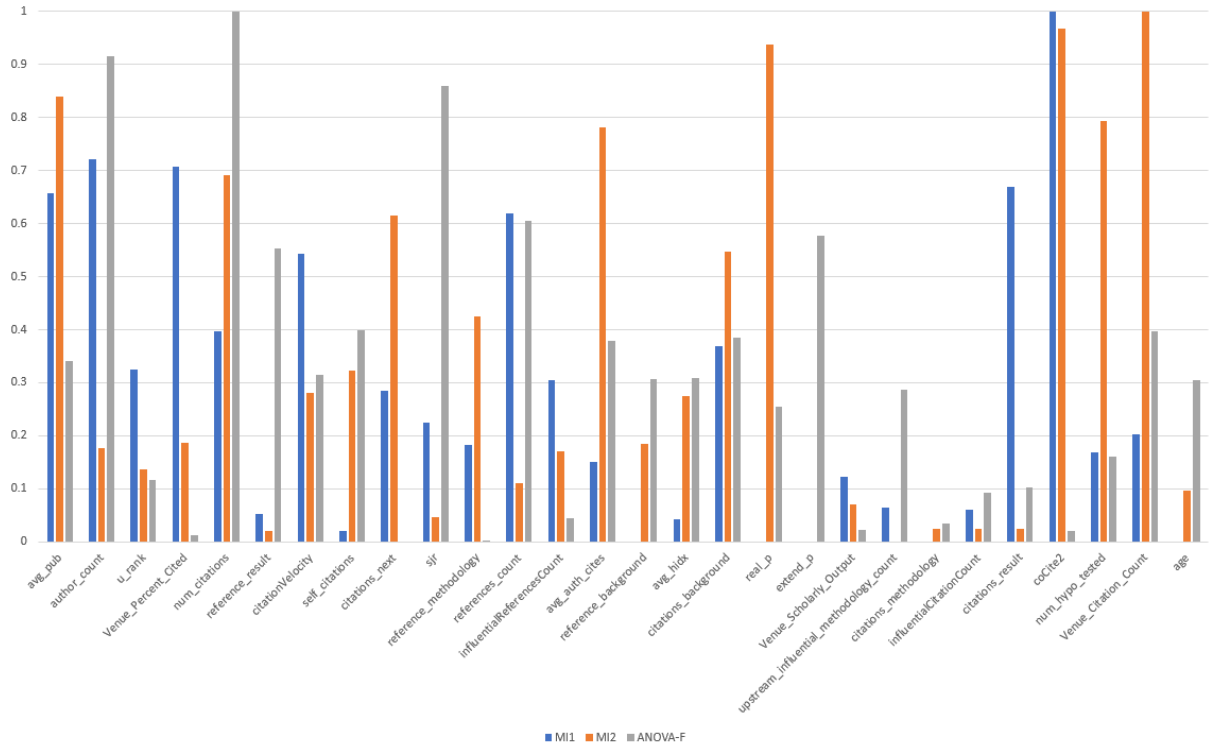


Figure 4.5: Distribution of mutual information using scikit-learn(MI1), mutual information from pyitlib(MI2) and ANOVA-F (normalized) values of *independent features*. The x-axis labels are feature IDs used in these calculations.

almost the best performance with the top 8 features identified using ANOVA-F and Mutual Information. Adding more features do not seem to help.

4.4.5 Selecting Top Features

As seen in Table 4.2, 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 uniform around each point. In our implementation[2], 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. We

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

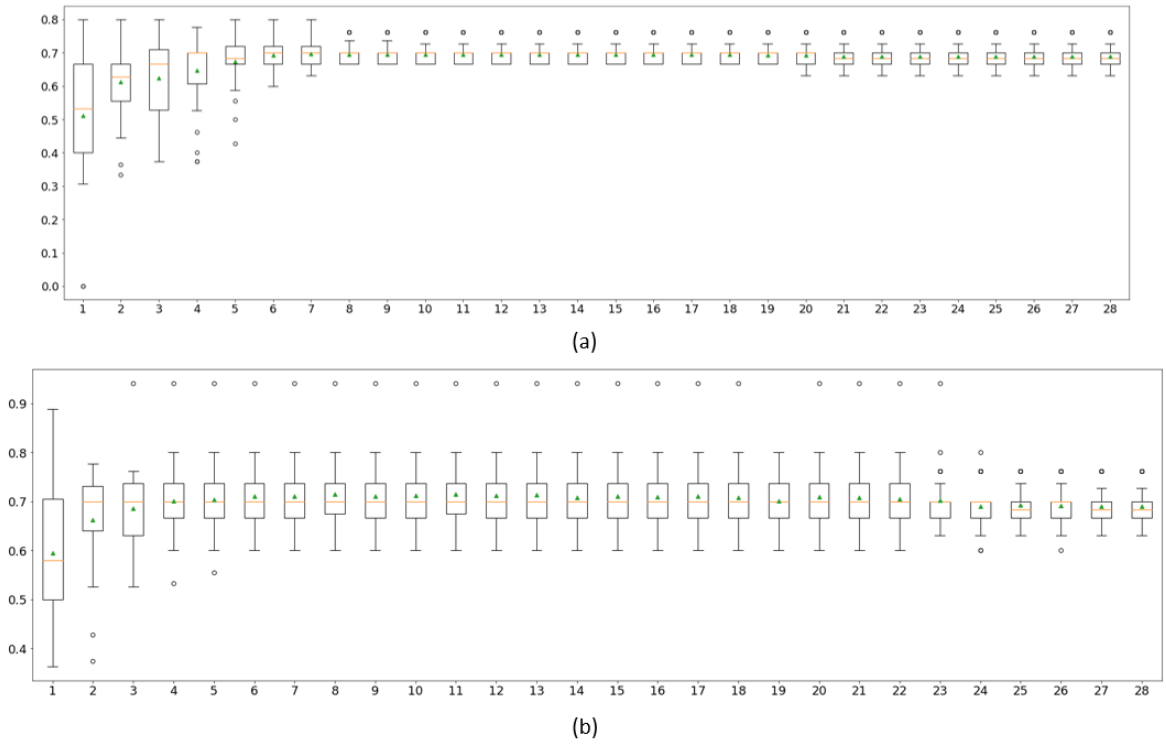


Figure 4.6: Box whisker plots of SVM F1 features for each number of selected features using (a) ANOVA-F and (b) Mutual information. The green triangles represent arithmetic means. The open dots are outliers beyond caps. The red short lines show medians.

select a set of 9 features, marked with a orange dot in Table 4.2, by our assesment of values from MI1, MI2 and ANOVA-F. These top 9 features are from different categories of feature sets like bibliometric(self_citations, influentialReferencesCount, influentialCitationCount), author(author_count, avg_auth_cites),statistical(num_hypo_tested) and semantic(citations_methodology,reference_result, upstream_influential_methodology_count).

4.4.6 Feature normalization

The goal of feature normalization is to change the values of the numeric columns on a common scale, without distorting the range of values. In some cases, due to the dissimilar ranges, gradients may end up taking a long time to find its way to the global/local minimum. We make sure that the different features take on similar ranges of values so that gradient descents can converge more quickly by applying feature normalization.

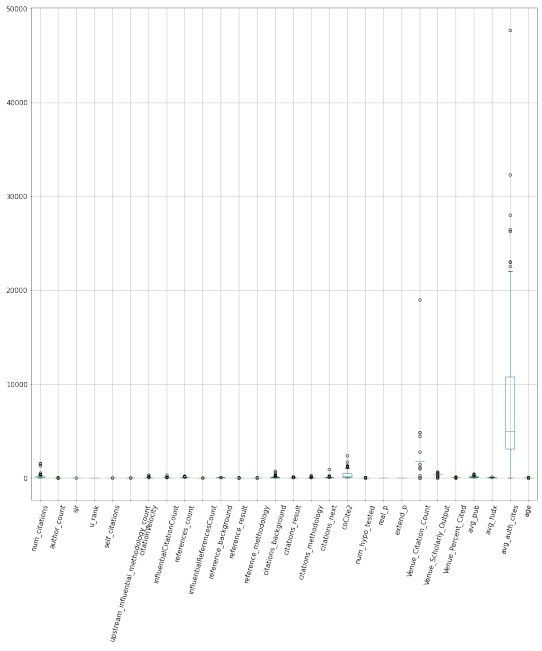
Table 4.2: Top 10 features identified by MI (top portion) and ANOVA-F (bottom portion). Feature IDs correspond to x-labels in Figure 4.5. We show their normalized MI and ANOVA-F values. Blue text are cross-listed features by both MI and ANOVA-F.

ID		Feature	MI1	MI2	ANOVA-F
22		Venue_citation_count	0	0.026	0.036
18		coCite2	0	0.67	0.255
28		age	0	0.076	0.305
15		citations_result	0	0.14	0.308
19		num_hypo_tested	0	0.06	0.578
8		influentialCitationCount	0.021	0.23	0.340
16	•	citations_methodology	0.043	0.13	0.310
6		upstream_influential_methodology_count	0.053	0.02	0.552
23		Venue_Scholarly_Output	0.062	0.02	0.093
21		extend_p	0.066	0.0	0.287
5	•	self_citations	0.397	0.50	1.00
2	•	author_count	0.721	0.13	0.915
10	•	influentialReferencesCount	0.224	0.04	0.860
12	•	reference_result	0.620	0.08	0.605
19	•	num_hypo_tested	0	0.06	0.578
6	•	upstream_influential_methodology_count	0.053	0.13	0.552
8	•	influentialCitationCount	0.021	0.23	0.340
27	•	avg_auth_cites	0.204	0.71	0.397
17		citations_next	0.370	0.39	0.385
14		citations_background	0.152	0.56	0.380

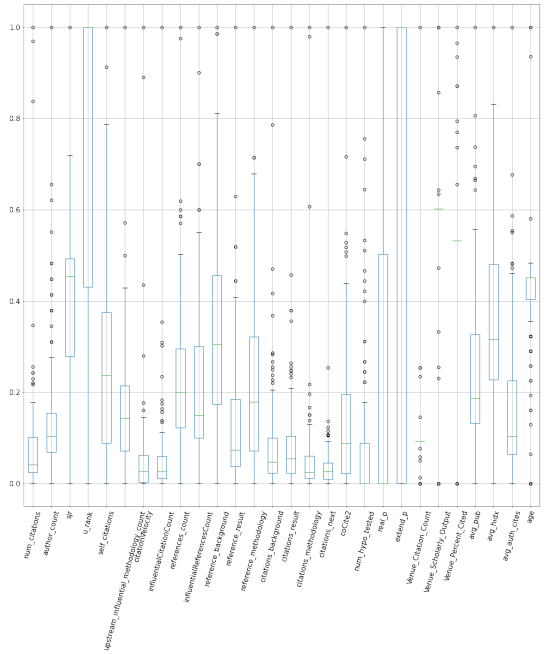
We use MinMaxScaler function [54] to transform features within a given range. The estimator translates each feature such that the training set is in the given range of $[0, 1]$.

4.5 Results

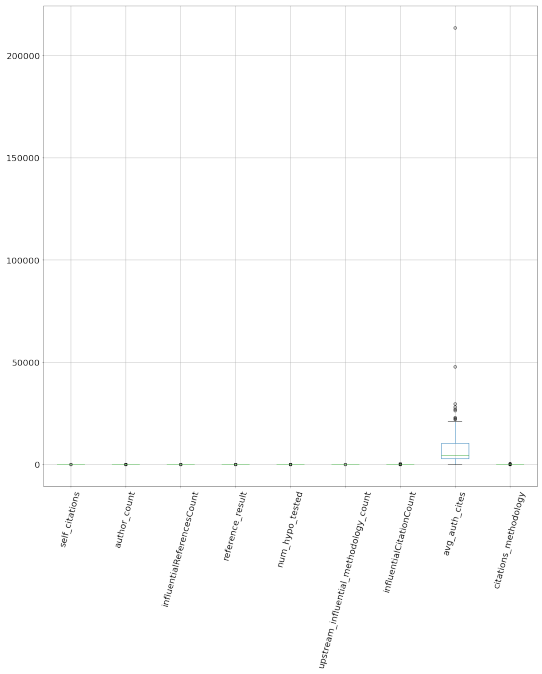
Because the sample size is relatively small, we apply a five-fold cross validation (CV) for all models described in 4.3. Figure 4.8 shows that evaluation results using the core features exhibit significantly different performances. The highest F_1=0.68 is achieved by SVM and QDA, followed by LR with 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.



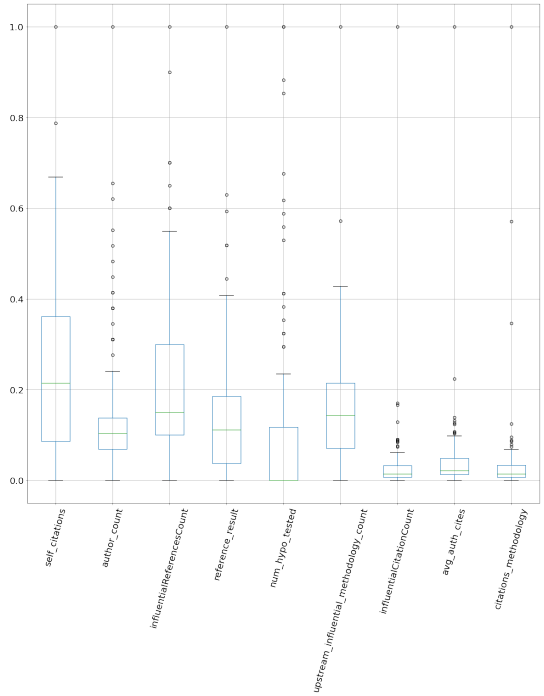
(a) Independent features



(b) Normalized independent features



(c) Top 9 features



(d) Normalized top 9 features

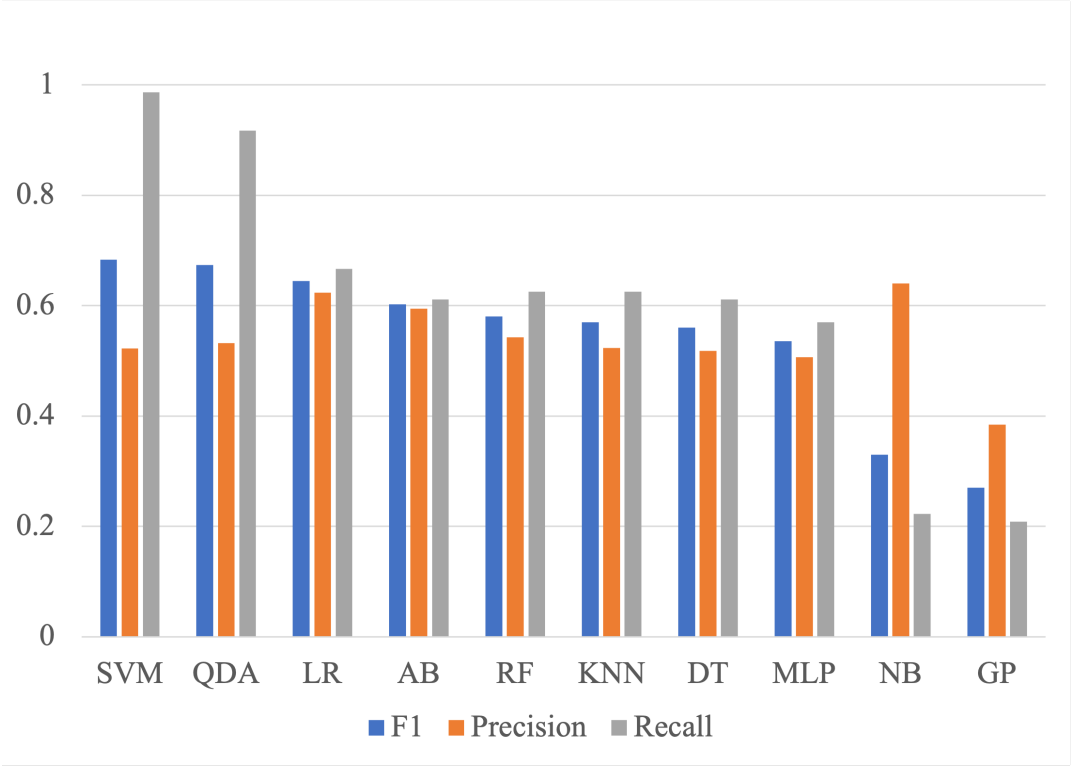


Figure 4.8: Five-fold CV results using core features sorted by F1 from left to right.

We compare the performance of the models when different sets of features are selected by methods described in 4.4. We observe consistent performance across other classifiers. This also helps verify top features identified above in 4.4.5. 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 independent 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 4.9.

The F1 of SVM is stable with a marginal decrease with independent and top features. Quadratic Discriminant Analysis(QDA), Logistic Regression(LR), AdaBoost(AB), K-nearest neighbors(K-NN), and Decision Trees(DT) exhibit a general decline when trained with independent features and top-9 features. Normalizing features helps boosting performances in certain conditions. In Random Forests(RF), Muti-layer Perceptron(MLP),

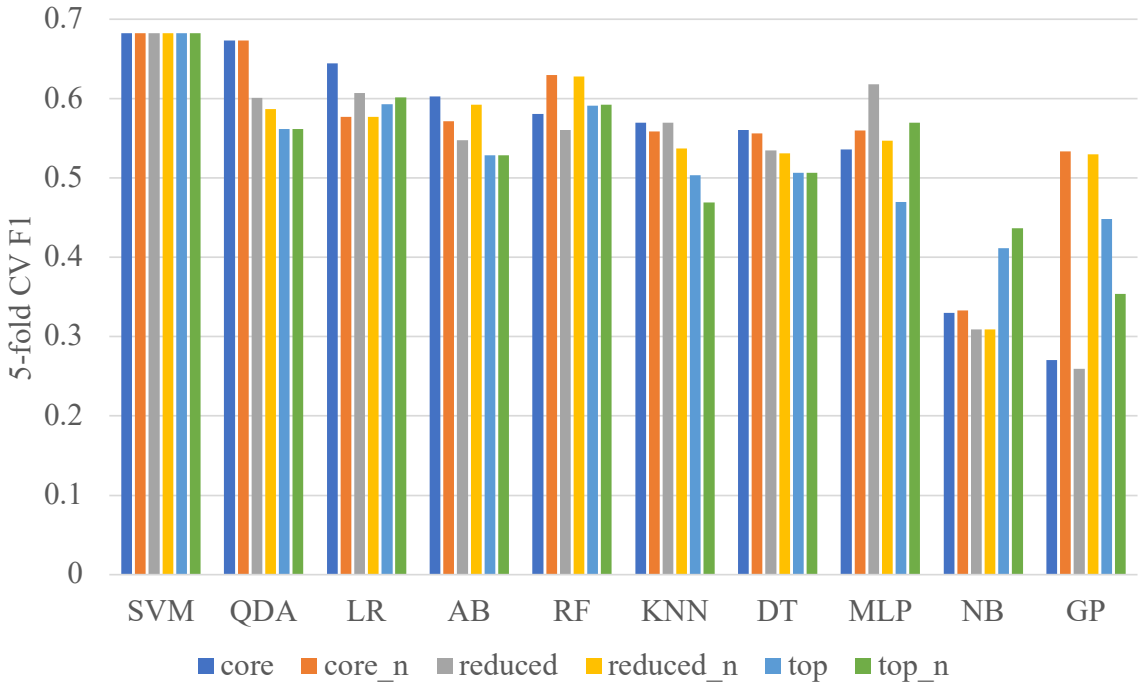


Figure 4.9: Comparison of F1-measures of classifiers trained on core features, normalized core features, reduced features, normalized reduced features, top 9 features, and normalized top 9 features.

Naive Bayes(NB), and Gaussian Process(GP), classifiers trained on the top-9 features outperform the core and independent features. Normalizing feature may or may not boost the performances.

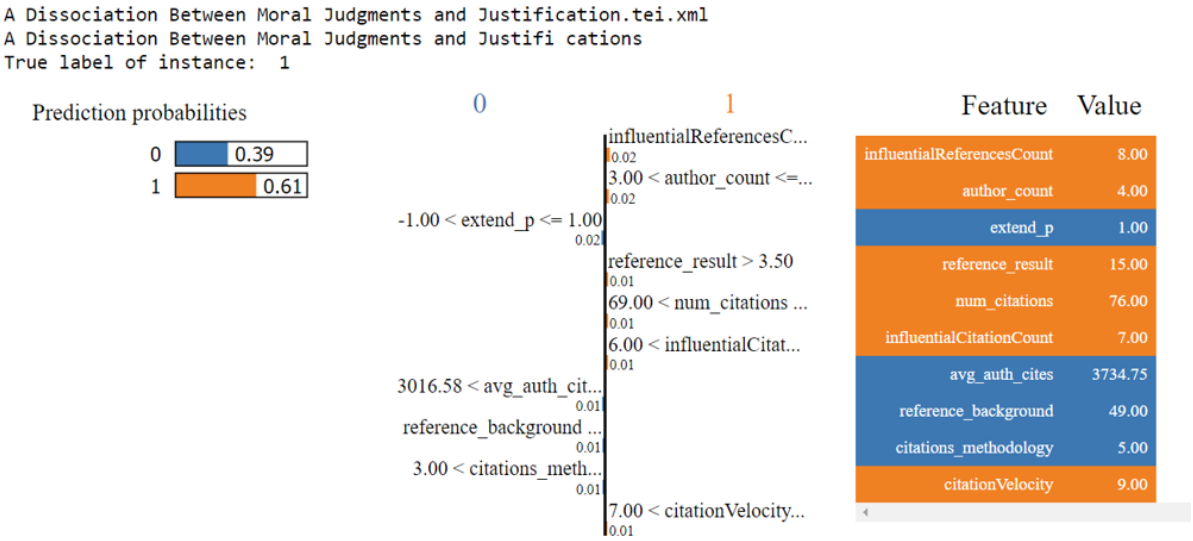
Except for QDA, which exhibits a drop of ~ 0.11 and KNN, which exhibits a drop of ~ 0.07, the other classifiers either show a mild decrease (< 0.05) or an increase between core features and top-9 features. The increase of F1 with top-9 features indicate that the classifier may overfit when trained with core features or independent features, which can be mitigated by adding more training samples. Overall, Figure 4.9 verifies that the top-9 features selected in Section 4.4.5 produce generally consistent results across most classifiers except for QDA and KNN. Feature normalization helps to boost performances in certain cases.

4.6 Interpretability

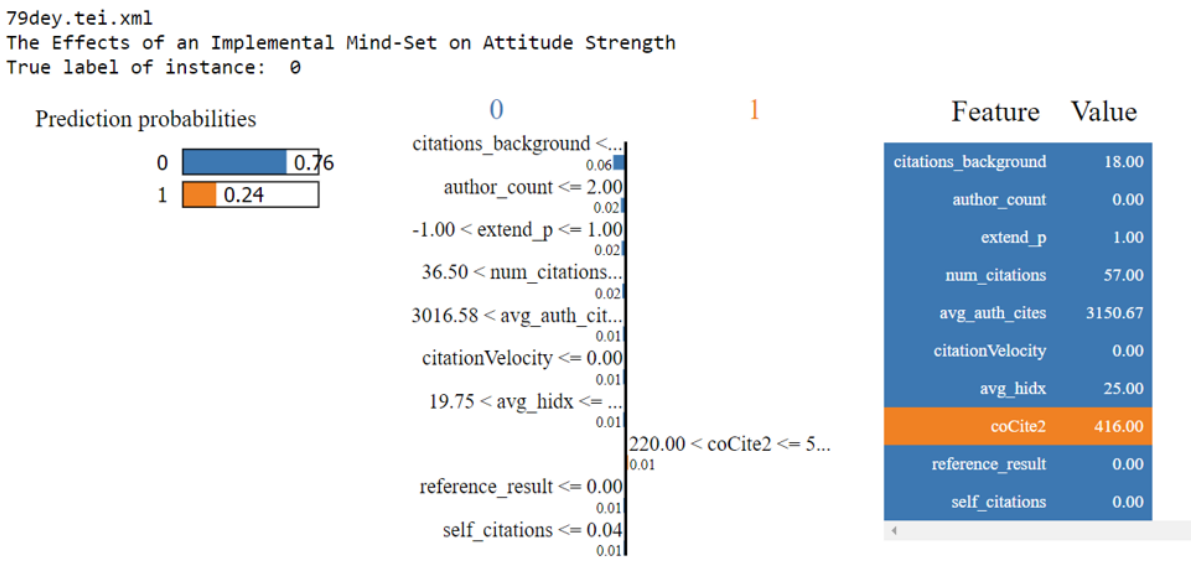
Interpretability is a technique used to explain the predictions made by a model. It is important to understand the mechanics of the technique. Through this thesis, we are trying to assess the feature importance of the different metrics of a scientific research article to predict its reproducibility. So far, we have used various feature selection techniques and observed the performance of supervised machine learning models on those sets of features. These give some idea of which features can be of importance. Further, we want to use a interpretable technique to investigate the model performance.

We use LIME - Local Interpretable Model-agnostic Explanations [21]. It explains the predictions of any classifier by learning an interpretable model locally around the prediction. The output of LIME is a list of explanations, reflecting the contribution of each feature to the prediction of a data sample. In the current implementation, linear models are used to approximate local behaviour.

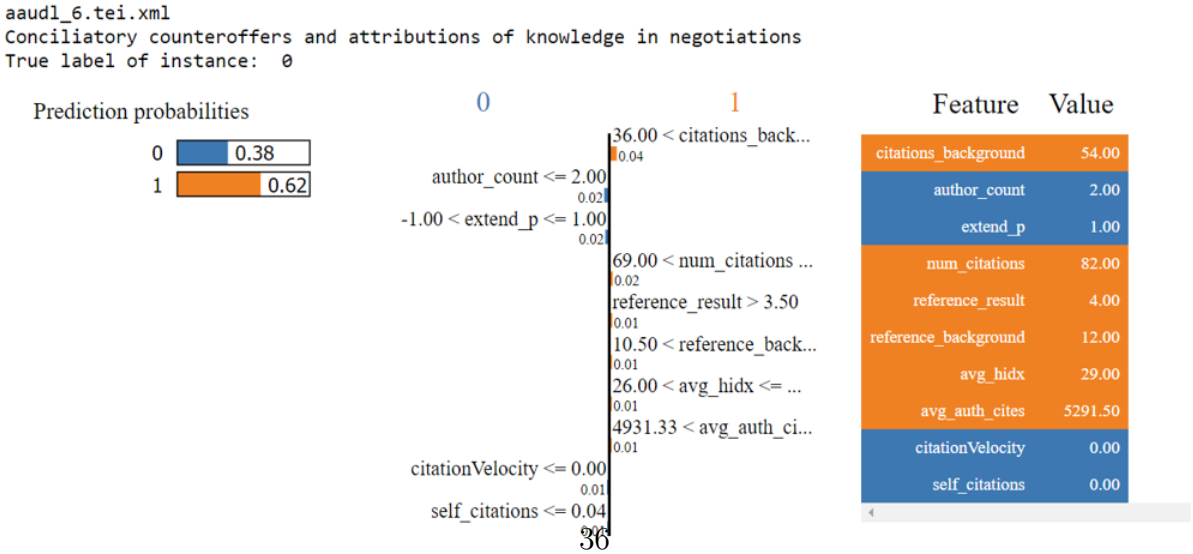
In our experiments, we use the random forests classifier with same parameters as in 4.3. It achieves an accuracy of 0.57 on the independent features. For the LIME explainer, statistics are computed for each feature. If the feature is numerical, the mean and standard deviation are used to discretize into quartiles. If the feature is categorical, the frequency of each value is used. The LimeTabularExplainer module explains classifiers that use tabular data. We pick samples from the test dataset randomly to explain that instance. Three examples are shown in figure 4.10. LIME explains the probability of random forests model prediction by the probabilities of each feature determining that prediction. We observe different categories of features appearing in each explanation. The features that appear in all three examples are: author_count, extend_p, reference_result, num_citations, avg_auth_cities and citation_velocity out of which 3 are from the top 9 features. In the first correctly predicted example, 67% of the influential features for correct prediction are from the top 9 features.



(a) Correctly predicted reproducible paper



(b) Correctly predicted non-reproducible paper



(c) Incorrectly predicted non-reproducible paper

Figure 4.10: LIME Examples

4.7 Discussion

Among the different machine learning models used, the Support Vector Machine(SVM) model gives the highest accuracy of 71% using the top 9 features. This is consistent with the results obtained by previous work described in Chapter 2: Black box statistical models to predict replicability(1) and Estimating deep replicability using human and artificial intelligence(2).

The black statistical models to predict replicability [28] method approximated accuracy of 70%. We used the the similar dataset for our work excluding Experimental Economics Replication Project(EERP). They chose to run the experiments on the random forests model because of its result of averaging over a "forest" of decision trees and ability to fit over random subset of variables to predict observations. We use the similar set of features including the statistical design properties(sample size, p-value, effect size, etc) and descriptive aspects(citations, author success, funding, etc.) of the original study. One set that we did not use was the characteristic differences between the original and replication experiments(conducted in the same country or same pool of subjects). It was observed that excluding the replication features did not affect the accuracy since these were standardized features. The statistical features such as p-value and effect size were of highest importance. This was followed by number of authors, maximum author citations, discipline, and citations. The accuracy increases when the nature of finding, number of authors, paper length and funding are added to the feature set. We have also included the author count and average author citations in our top 9 features.

The work on estimating deep replicability using human and artificial intelligence [29] predicted replication with accuracy in the range 0.65 to 0.78. The authors observed that the papers that failed to replicate from the Replication Project [25] were cited at the same yearly rate as papers that were replicated. Thus, we excluded number of citations from the top 9 features and included the upstream citations of the paper as well as the

papers influential citations instead. Their model based on linguistic features did not detect the importance of authorship, prestige, sex of authors, discipline, journal specific words or subjective probabilities/persuasive language. Similar features from our set like author h-index, subject, subject code weren't used in the top 9 features. They used the prediction market or survey predictions in combination with the machine learning methods which performed better than any other method. Text is shown to provide more consistent and higher predictability of the reviewer metrics. This can be attempted in the future work to understand the effect of textual features in addition to our experiments.

Chapter 5
Conclusion

Through the findings in this thesis, we can trust the influence machine learning models can have on the prediction of reproducibility of research articles. We observe promising output from a relatively small dataset with a variety of features for each paper. By conducting statistical correlation and feature analysis, we were able to identify 9 top features, which were believed 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. The interpretability of the model helps understand the features selected for a particular prediction instance. Though it is difficult to conclusively determine the perfect feature set to predict reproducibility, our work delivers a preliminary study in assessing the feature importance for the task.

Our study has three 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., [59]. With the advocacy and adoption of open science, more papers will be labeled by domain experts, e.g., the repliCATS project [60, 61], 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.

The feature extraction framework directly retrieves numerical values from the Scopus, CrossRef and Semantic Scholar(S2) APIs. These values may themselves be extracted from machine learning models with their own inconsistencies. These will be carried on to our framework and experiments. Designing our own models for this purpose can be non-trivial.

Bibliography

[1] Fidler, F. and J. Wilcox (2019) "Reproducibility of Scientific Results," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Winter 2019 ed., Metaphysics Research Lab, Stanford University.

[2] Leek, J. T. and L. R. Jager (2017) "Is Most Published Research Really False?" *Annual Review of Statistics and Its Application*, **4**(1), pp. 109–122, https://doi.org/10.1146/annurev-statistics-060116-054104.
URL https://doi.org/10.1146/annurev-statistics-060116-054104

[3] (2015) "Estimating the reproducibility of psychological science," *Science*, **349**(6251), https://science.sciencemag.org/content/349/6251/aac4716.full.pdf.
URL https://science.sciencemag.org/content/349/6251/aac4716

[4] Prinz, F., T. Schlange, and K. Asadullah (2011) "Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov 10: 712," *Nature reviews. Drug discovery*, **10**, p. 712.

[5] Begley, C. and L. Ellis (2012) "Drug development: raise standards for preclinical cancer research. Nature 483(7391):531-533, Epub 2012/03/31," *Nature*, **483**, pp. 531–3.

[6] Freedman, L., I. Cockburn, and T. Simcoe (2015) "The Economics of Reproducibility in Preclinical Research," *PLoS biology*, **13**, p. e1002165.

[7] Ioannidis, J. and C. Doucouliagos (2013) "What's to Know About the Credibility of Empirical Economics?: Scientific Credibility of Economics," *Journal of Economic Surveys*, **27**.

[8] Maniadis, Z., F. Tufano, and J. A. List (2014) "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects," *The American Economic Review*, **104**(1), pp. 277–290.
URL http://www.jstor.org/stable/42920695

[9] Camerer, C., A. Dreber, E. Forsell, T. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen,

and H. Wu (2016) "Evaluating replicability of laboratory experiments in economics," *Science*, **351**.

[10] Button, K., J. Ioannidis, C. Mokrysz, B. Nosek, J. Flint, E. Robinson, and M. Munafò (2013) "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience," *Nature reviews. Neuroscience*, **14**.

[11] Simmons, J., L. Nelson, and U. Simonsohn (2011) "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant," *Psychological Science*, **20**, pp. 1–8.

[12] Carpenter, S. (2012) "Psychology's Bold Initiative," *Science (New York, N.Y.)*, **335**, pp. 1558–61.

[13] Fiedler, S. and O. Collaboration (2012) "An open, large-scale, collaborative effort to estimate the reproducibility of psychological science." *Perspectives on Psychological Science*, **7**, pp. 657–660.

[14] Bohannon, J. (2014) "Replication effort provokes praise - And 'bullying' charges," *Science*, **344**, pp. 788–789.

[15] Begley CG, I. J. (2015) "Reproducibility in science," *Circulation Research*, pp. 116–126.

[16] Baker, M. (2016) "1,500 scientists lift the lid on reproducibility," *Nature*, **533**, pp. 452–454.

[17] Klein, R., K. Ratliff, M. Vianello, R. Jr, Bahník, M. Bernstein, K. Bocian, M. Brandt, B. Brooks, C. Brumbaugh, Z. Cemalcilar, J. Chandler, W. Cheong, W. Davis, T. Devos, M. Eisner, N. Frankowska, D. Furrow, E. Galliani, and B. Nosek (2013) "Investigating Variation in Replicability A "Many Labs" Replication Project," *Social Psychology*, **45**.

[18] Klein, R. A., M. Vianello, F. Hasselman, B. G. Adams, J. Reginald B. Adams, S. Alper, M. Aveyard, J. R. Axt, M. T. Babalola, Štěpán Bahník, R. Batra, M. Berkics, M. J. Bernstein, D. R. Berry, O. Bialobrzeska, E. D. Binan, K. Bocian, M. J. Brandt, R. Busching, A. C. Rédei, H. Cai, F. Cambier, K. Cantarero, C. L. Carmichael, F. Ceric, J. Chandler, J.-H. Chang, A. Chatard, E. E. Chen, W. Cheong, D. C. Cicero, S. Coen, J. A. Coleman, B. Collisson, M. A. Conway, K. S. Corker, P. G. Curran, F. Cushman, Z. K. Dagona, I. Dalgar, A. D. Rosa, W. E. Davis, M. de Bruijn, L. D. Schutter, T. Devos, M. de Vries, C. Doğulu, N. Dozo, K. N. Dukes, Y. Dunham, K. Durrheim, C. R. Ebersole, J. E. Edlund, A. Eller, A. S. English, C. Finck, N. Frankowska, M. Ángel Freyre, M. Friedman, E. M. Galliani, J. C. Gandi, T. Ghoshal, S. R. Giessner, T. Gill, T. Gnambs, Ángel Gómez, R. González, J. Graham, J. E. Grahe, I. Grahek, E. G. T. Green, K. Hai, M. Haigh, E. L. Haines, M. P. Hall,

M. E. Heffernan, J. A. Hicks, P. Houdek, J. R. Huntsinger, H. P. Huynh, H. IJzerman, Y. Inbar, Åse H. Innes-Ker, W. Jiménez-Leal, M.-S. John, J. A. Joy-Gaba, R. G. Kamiloğlu, H. B. Kappes, S. Karabati, H. Karick, V. N. Keller, A. Kende, N. Kervyn, G. Knežević, C. Kovacs, L. E. Krueger, G. Kurapov, J. Kurtz, D. Lakens, L. B. Lazarević, C. A. Levitan, J. Neil A. Lewis, S. Lins, N. P. Lipsey, J. E. Losee, E. Maassen, A. T. Maitner, W. Malingumu, R. K. Mallett, S. A. Marotta, J. Međedović, F. Mena-Pacheco, T. L. Milfont, W. L. Morris, S. C. Murphy, A. Myachykov, N. Neave, K. Neijenhuijs, A. J. Nelson, F. Neto, A. L. Nichols, A. Ocampo, S. L. O'Donnell, H. Oikawa, M. Oikawa, E. Ong, G. Orosz, M. Osowiecka, G. Packard, R. Pérez-Sánchez, B. Petrović, R. Pilati, B. Pinter, L. Podesta, G. Pogge, M. M. H. Pollmann, A. M. Rutchick, P. Saavedra, A. K. Saeri, E. Salomon, K. Schmidt, F. D. Schönbrodt, M. B. Sekerdej, D. Sirlopú, J. L. M. Skorinko, M. A. Smith, V. Smith-Castro, K. C. H. J. Smolders, A. Sobkow, W. Sowden, P. Spachtholz, M. Srivastava, T. G. Steiner, J. Stouten, C. N. H. Street, O. K. Sundfelt, S. Szeto, E. Szumowska, A. C. W. Tang, N. Tanzer, M. J. Tear, J. Theriault, M. Thomae, D. Torres, J. Traczyk, J. M. Tybur, A. Ujhelyi, R. C. M. van Aert, M. A. L. M. van Assen, M. van der Hulst, P. A. M. van Lange, A. E. van 't Veer, A. Vásquez-Echeverría, L. A. Vaughn, A. Vázquez, L. D. Vega, C. Verniers, M. Verschoor, I. P. J. Voermans, M. A. Vranka, C. Welch, A. L. Wichman, L. A. Williams, M. Wood, J. A. Woodzicka, M. K. Wronska, L. Young, J. M. Zelenski, Z. Zhijia, and B. A. Nosek (2018) "Many Labs 2: Investigating Variation in Replicability Across Samples and Settings," *Advances in Methods and Practices in Psychological Science*, **1**(4), pp. 443–490, https://doi.org/10.1177/2515245918810225. URL https://doi.org/10.1177/2515245918810225

[19] Camerer, C., A. Dreber, F. Holzmeister, T. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. Nosek, T. Pfeiffer, A. Altmejd, N. Buttrick, T. Chan, Y. Chen, E. Forsell, A. Gampa, E. Heikensten, L. Hummer, T. Imai, and H. Wu (2018) "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015," *Nature Human Behaviour*, **2**.

[20] Laraway, S., S. Snycerski, S. Pradhan, and B. Huitema (2019) "An Overview of Scientific Reproducibility: Consideration of Relevant Issues for Behavior Science/-Analysis," *Perspectives on Behavior Science*, **42**, pp. 1–25.

[21] Ribeiro, M. T., S. Singh, and C. Guestrin (2016) ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[22] Munafò, M., B. Nosek, D. Bishop, K. Button, C. Chambers, N. Percie du Sert, U. Simonsohn, E.-J. Wagenmakers, J. Ware, and J. Ioannidis (2017) "A manifesto for reproducible science," *Nature Human Behaviour*, **1**, p. 0021.

[23] Ioannidis, J., D. Fanelli, D. Dunne, and S. Goodman (2015) "Meta-research: Evaluation and Improvement of Research Methods and Practices," *PLOS Biology*, **13**, p. e1002264.

[24] Szűcs, D. (2016) "Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature," .

[25] Nosek, B., T. Hardwicke, H. Moshontz, A. Allard, K. Corker, A. Dreber, F. Fidler, J. Hilgard, M. Kline, M. Nuijten, J. Rohrer, F. Romero, A. Scheel, L. Scherer, F. Schönbrodt, and S. Vazire (2021) "Replicability, Robustness, and Reproducibility in Psychological Science," .

[26] Dreber, A., T. Pfeiffer, J. Almenberg, S. Isaksson, B. Wilson, Y. Chen, B. Nosek, and M. Johannesson (2015) "Using prediction markets to estimate the reproducibility of scientific research," *Proceedings of the National Academy of Sciences of the United States of America*, **112**.

[27] Gelman, B., C. Clark, S. E. Friedman, U. Kuter, and J. E. Gentile (2021) "Toward A Robust Method for Understanding the Replicability of Research." in *SDU@ AAAI*.

[28] Altmejd, A., A. Dreber, E. Forsell, J. Huber, T. Imai, M. Johannesson, M. Kirchler, G. Nave, and C. Camerer (2019) "Predicting the replicability of social science lab experiments," *PLOS ONE*, **14**, p. e0225826.

[29] Yang, Y., 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**, p. 201909046.

[30] Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean (2013) "Distributed Representations of Words and Phrases and their Compositionality," *Advances in Neural Information Processing Systems*, **26**.

[31] Pawel, S. and L. Held (2020) "Probabilistic forecasting of replication studies," *PLoS ONE*, **15**.

[32] Wu, J., R. Nivargi, S. S. T. Lanka, A. Menon, S. A. Modukuri, N. Nakshatri, X. Wei, Z. Wang, J. Caverlee, S. Rajtmajer, and C. L. Giles (2021) "Predicting the Reproducibility of Social and Behavioral Science Papers Using Supervised Learning Models," *ArXiv*, **abs/2104.04580**.

[33] John, L., G. Loewenstein, and D. Prelec (2012) "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling," *Psychological Science*, **23**, pp. 524 – 532.

[34] Simmons, J., L. Nelson, and U. Simonsohn (2011) "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant," *Psychological Science*, **20**, pp. 1–8.

[35] Lopez, P. (2009) "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, Springer-Verlag, Berlin, Heidelberg, pp. 473–474.

[36] Tkaczyk, D., P. Szostek, M. Fedoryszak, P. J. Dendek, and Ł. Bolikowski (2015) "CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature," *IJDAR*, **18**(4), pp. 317–335.

[37] Wu, J., J. Killian, H. Yang, K. Williams, S. R. Choudhury, S. Tuarob, C. Caragea, and C. L. Giles (2015) "PDFMEF: A Multi-Entity Knowledge Extraction Framework for Scholarly Documents and Semantic Search," in *Proceedings of the 8th International Conference on Knowledge Capture*.

[38] Lipinski, M., K. Yao, C. Breitinger, J. Beel, and B. Gipp (2013) "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, pp. 385–386.

[39] Tkaczyk, D., A. Collins, P. Sheridan, and J. Beel (2018) "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, pp. 99–108.

[40] Kirkpatrick, K. (2016) "Search Engine's Author Profiles Now Driven By Influence Metrics," *Communications of ACM*.
URL `https://cacm.acm.org/news/201387-search-engines-author-profiles-now-driven-fulltext`

[41] Aksnes, D., L. Langfeldt, and P. Wouters (2019) "Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories," *SAGE Open*, **9**, p. 215824401982957.

[42] Valenzuela, M., V. A. Ha, and O. Etzioni (2015) "Identifying Meaningful Citations," in *AAAI Workshop: Scholarly Big Data*.

[43] Valenzuela, M., V. Ha, and O. Etzioni (2015) "Identifying meaningful citations," in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[44] Seeber, M., M. Cattaneo, M. Meoli, and P. Malighetti (2017) "Self-citations as strategic response to the use of metrics for career decisions," *Research Policy*, **48**, pp. 478–491.

[45] Kacem, A., J. W. Flatt, and P. Mayr (2020) "Tracking self-citations in academic publishing," *Scientometrics*, **123**, pp. 1157–1165.

[46] Petersen, A., S. Fortunato, R. K. Pan, K. Kaski, O. Penner, M. Riccaboni, H. Stanley, and F. Pammolli (2013) "Reputation and Impact in Academic Careers," *Proceedings of the National Academy of Sciences*.

[47] Sathian, B., P. Simkhada, and J. Sreedharan (2014) "Importance of the authorship order and number of co authors in a Publication for Evaluation: A Necessary Enquiry," *Nepal Journal of Epidemiology*, **4**, pp. 384–85.

[48] Lewallen, L. P. and P. B. Crane (2010) "Choosing a Publication Venue," *Journal of Professional Nursing*, **26**(4), pp. 250–254.
URL https://www.sciencedirect.com/science/article/pii/S8755722309001847

[49] Teixeira da Silva, J. A. (2020) "CiteScore: Advances, Evolution, Applications, and Limitations," *Publishing Research Quarterly*, **36**(3), pp. 459–468.

[50] Leydesdorff, L. and T. Opthof (2010) "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), pp. 2365–2369.

[51] Bast, H. and C. Korzen (2017) "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*, IEEE Computer Society, pp. 99–108.

[52] Jurgens, D., S. Kumar, R. Hoover, D. A. McFarland, and D. Jurafsky (2018) "Measuring the Evolution of a Scientific Field through Citation Frames," *TACL*, **6**, pp. 391–406.

[53] Wu, J., P. Wang, X. Wei, S. Rajtmajer, C. L. Giles, and C. Christopher (2020) "Acknowledgement Entity Recognition in CORD-19 Papers," in *Proceedings of the 1st Workshop on Scholarly Document Processing*.

[54] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011) "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, **12**, pp. 2825–2830.

[55] Kendall, M. G. (1938) "A New Measure of Rank Correlation," *Biometrika*, **30**(1-2), pp. 81–93.
URL https://doi.org/10.1093/biomet/30.1-2.81

[56] Kraskov, A., H. Stögbauer, and P. Grassberger (2004) "Estimating mutual information," *Phys. Rev. E*, **69**, p. 066138.

[57] Ross, B. C. (2014) "Mutual Information between Discrete and Continuous Data Sets," *PLOS ONE*, **9**(2), pp. 1–5.
URL https://doi.org/10.1371/journal.pone.0087357

[58] Kuhn, M. and K. Johnson (2019) *Feature Engineering and Selection: : A Practical Approach for Predictive Models*, Chapman and Hall/CRC.

[59] Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, and H. Wu (2016) "Evaluating replicability of laboratory experiments in economics," *Science*, **351**(6280), pp. 1433–1436.

[60] Hanea, A., D. P. Wilkinson, M. McBride, A. Lyon, D. van Ravenzwaaij, F. Singleton Thorn, C. T. Gray, D. R. Mandel, A. Willcox, E. Gould, and et al. (2021) "Mathematically aggregating experts' predictions of possible futures," .

[61] Nosek, B. A., T. E. Hardwicke, H. Moshontz, A. Allard, K. S. Corker, A. Dreber, F. Fidler, J. Hilgard, M. Kline Struhl, M. B. Nuijten, and et al. (2021) "Replicability, Robustness, and Reproducibility in Psychological Science," .
URL psyarxiv.com/ksfvq