The Pennsylvania State University

The Graduate School

UNDERSTANDING AND PREDICTING RETRACTIONS OF PUBLISHED WORK

A Thesis in

Computer Science and Engineering

by

Sai Ajay Modukuri

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science

August 2021

The thesis of Sai Ajay Modukuri was reviewed and approved by the following:

Clyde Lee Giles
David Reese Professor of Information Sciences and Technology
Thesis Co-Advisor

Sarah Rajtmajer
Associate Professor of Information Sciences and Technology
Thesis Co-Advisor

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

Recent increases in the number of retractions of published papers reflect heightened attention and increased scrutiny in the scientific process motivated, in part, by the replication crisis. These trends motivate computational tools for understanding and assessment of the scholarly record. Here, we sketch the landscape of retracted papers in the Retraction Watch database, a collection of 19k records of published scholarly articles that have been retracted for various reasons (e.g., plagiarism, data error). Using metadata as well as features derived from full-text for a subset of retracted papers in the social and behavioral sciences, we develop a random forest classifier to predict retraction in new samples with 73% accuracy and F1-score of 71%. We believe this study to be the first of its kind to demonstrate the utility of machine learning as a tool for the assessment of retracted work.

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. I would like to specialy thank Professor Sarah Rajtmajer for this opportunity and the support. It has been an great learning experience. I would like to thank my advisors, Prof. Lee Giles, Prof. Sarah Rajtmajer, Prof. Anna Cinzia Squicciarini and Prof. Jian Wu for their support and guidance throughout my time with the team. I would like to acknowledge the contributions of my team members: Arjun Manoj Menon, Sree Sai Teja Lanka and Rajal for helping me extract and develop features. Special thanks to my committee member Rui Zhang for participating in my defense.

This work was partially supported by the Defense Advanced Research Projects Agency cooperative agreement No. W911NF-19-2- 0272. The ideas in this paper do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

Chapter 1
Introduction

The last two decades have seen growing concern in the scientific community about the integrity of published work [1–3] and an increase in the number of retractions of published articles (see Figure 2.1), in part due to increased scrutiny and improved oversight [4–6]. Focused studies of the primary reasons for retraction have suggested that research misconduct and fraud make up the majority, but also that a sizeable number of retractions are due to laboratory error, error in analyses, or inability to submit to reproduction or replication [7, 8].

Continued attention to and assessment of our confidence in published work is the cornerstone to efficient scientific progress, while the sheer volume of research papers published each year is overwhelming and increasing [9]. Auditors and stakeholders, including reviewers, editors, other scientists, and the broader public, seek indicators and tools to contextualize and evaluate published findings, but these processes are still largely ad hoc. Proxies for credibility, such as citations and impact factors, while widespread, have also been shown to be biassed and flawed [10–12]. Leading voices have argued for a re-imagining of scholarship itself [13, 14] in support of greater transparency and verifiability. While it is still unclear what form they must take, it is clear that computational tools will play a role in aggregating, sorting, querying, and evaluating scientific outputs in the future. Our work is motivated by this view, as we put forward a

supervised approach to determine factors that best predict the retraction of scholarly work.

The next section highlights related work in the area of understanding the retraction of scientific publications. Section 3 sketches our primary dataset and preprocessing pipeline. Section 4 outlines our features pulled from metadata and full-text documents. Sections 5 and 6 detail our classification approach and ablation studies. We conclude with a discussion of our findings and implications for ongoing and future work.

1.1 Background

1.2 Motivation

Here, we study retractions collected by The Center for Scientific Integrity and included in its Retraction Watch database (retractionwatch.com; [15]). We extract a combination of metadata and full-text features that can separate retracted from non-retracted papers and develop a classifier to predict retraction in new samples with relatively high confidence. We focus on research publications in the social and behavioral sciences in this study, as it is not yet clear whether and how different research cultures and publishing norms differentially impact retraction across fields.

- Extract meaningful information from retracted papers in the social and behavioral sciences as well as from a complement set of non-retracted papers, from both metadata and full-text.
- Build a binary classifier to identify the likelihood of a paper's retraction given extracted information with 73% accuracy.
- Identify by ablation studies features and sets of features that best separate retracted from non-retracted papers. These insights, we argue, can direct further research into automated tools for assigning confidence in publication claims.

1.3 Approach

In this work we carefully analyze and understand retractions. Various insights into retractions and other relevant analyses to gain a better understanding. We then explore various features that can be reliably extracted through our extraction framework. We define various features and how they are calculated and extracted using this pipeline. We give explanation of the intuition of why we used a certain feature and our reasoning behind the feature used.

We then run tight analysis of how random forests can fit our extracted features and can reliably predict retractions. We furthermore perform exhaustive ablation studies to understand how various features contribute to better predictions of retracted papers.

Chapter 2
Related work

The Center for Open Science(COS), a non-profit organization which aims at improving openness, integrity and reproduciblity of scientific work, has launched a large scale effort to reproduce 100 published experimental and corelational studies [1] through its Open Science Framework(OSF) [16]. The study was conducted on articles from three leading psychology journals: Journal of Personality and Social Psychology (JPSP), Psychological Science (PSCI), and Journal of Experimental Psychology: Learning, Memory, and Cognition (JEP:LMC). This study, known as reproducability project [1] found that the mean of replication effect sizes ($M_r = 0.197$, $SD = 0.257$) were half the mean effect sizes of the original studies ($M_r = 0.403$, $SD = 0.188$). Even more concerning is that only 36% of the replications had statically significant results($p<0.5$), meaning only one-third of the experiment were replicable. One keen observation in this study is that replications varied widely accross journals: JPSP- 23%, JEP:LMC- 48%, PSCI- 82%.

Many Labs 2 [3] is another major replication project. It conducted replication studies on 28 classic published results. This replication effort involved $15,305$ participants from across 36 countries. With a statistical significance of $p < 0.5$ as criteria, the study was able to successfully replicate only 15 of the total 28 published results (54% successful replications). In a similar effort [2] to replicate 21 social scientific experiments, only 62% of the replicated studies produced effect size in the same direction of the original work

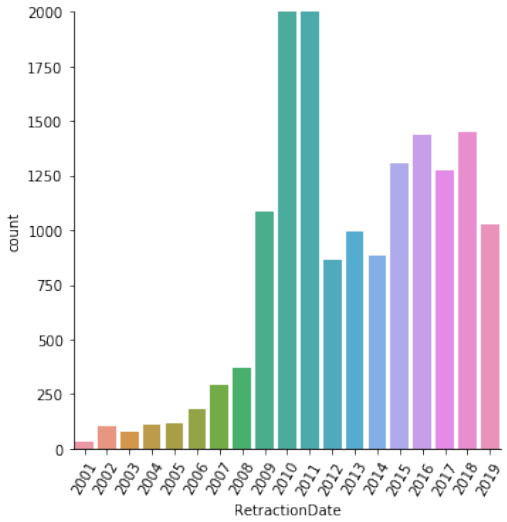

Figure 2.1: Increasing uptrend of retractions over the years 2001-2019 with the year of retraction on X-axis and number of retractions on Y-axis.

Incresing retractions

and the effect size of replication (on average) is only 50% of the original effect size. Many other replication studies [3, 17, 18] have failed to replicate. This issue is not just isolated to psychology. Replication failures can be found across other various subjects such as cancer biology [19, 20], economics [21], etc. These concerns are further compounded with the increase in number of retractions in the recent years.

While replications vet a published paper in terms of methods, data collection, results, etc., retractions cover a much wider spectrum of quality checks including fraud, misconduct, plagiarisms, failure to replications, etc. Although most of the retractions are due to misconduct or fraud [7, 8], a sizeable number of retractions are due to irreproducable results [7]. The number of retractions have been increasing over the past few years, refer Fig. 2.1. The recent study on retractions [5] attributes increase in retractions to improved scrutiny of published work and not necessarily an increase fraud or misconduct.

Several studies have explored the retracted literature within a specific field of interest. [22] analyses retracted papers in the obstetrics literature using the Retraction Watch database and PubMed. They present a breakdown of various metrics in that dataset,

including journal impact factors, reasons for retractions, number of citations received, h-index of authors, and type of articles. Other authors have engaged in similar discussions across a variety of fields, including chemistry and material Science [23], biomedical sciences [24], dentistry [25] and oncology [26].

One recent paper [27] surveys publication rates after the first retraction for biomedical researchers with multiple retracted publications. The study finds that publication rates of authors with multiple retractions, most of whom were associated with scientific misconduct, declined rapidly after their first retraction, but a small minority continued to publish regularly. Similarly, [28, 29] also found a decline in number of citations after retraction.

Other work supplements data-driven findings from the analysis of retracted papers in the literature with suggestions for the community. Authors of [30] highlight so-called continued influence effects, or the tendency of false beliefs to persist after correction and retraction, supporting their discussion through analysis of citations of retracted papers in downstream research articles. Their work puts forward a set of best practices for science communication scholars and practitioners. While, [31] analyses retractions due to conflict of interest and argues for greater transparency on the part of both journals and authors in disclosing financial interests.

More closely related to our work, two very recent papers have begun to suggest possible indicators of low credibility work. [32] suggests that Benford's law can be used to differentiate retracted academic papers that have employed fraudulent/manipulated data from other academic papers that have not been retracted. Specifically, the authors construct several Benford conformity measures based on the first significant digits contained in the articles and show deviation for 37 papers containing known academic fraud. Supporting a broader conversation about open science and the role of transparency in scientific processes, [33] study retraction rates in work with associated shared datasets. Authors found that published work with open data has fewer retractions, signaling higher

credibility.

Finally, with the recent outbreak of COVID-19 (SARS-CoV-2) and a flurry of scientific output related to the pandemic, the scientific community has also faced a surge in the number of retractions in publications related to COVID-19. Work done by [34], [35] studies retractions related to COVID-19 and highlight the need for better scrutiny of published papers.

Overall, there is an increased effort to vet published work both through retractions and replications. However, the sheer number of published papers makes it a herculean tasks for humans to manually scrutinize the quality of a each paper, suggesting the need for computational tools.

In order to aid the ongoing efforts, we analyze the largest database of retractions: Retraction Watch [15]. Using the analysis, we select a combination of metadata and full-text features which can distinguish good quality papers. The analytical reasoning for feature selection is further reasoned by surveying domain experts. We further this understating of which features contribute to quality papers by building a classifier to classify which papers are likely to be retracted.

Chapter 3
Dataset

At the time of writing, the Retraction Watch database [15] has 19,864 records of retracted papers. Our analysis considered 18,970 records in the dataset from the year 2001 to 2019. We further downselected 8,087 retractions in the social sciences for classification. Specifically, our classification task considered papers tagged by the Retraction Watch organization relating to the following subjects: Health Sciences (HSC, 5,396 papers), Social Sciences (SOC, 2,651 papers), and Humanities (HUM, 366 papers). More than one subject may be listed for a given paper.

Each record in the database includes a rich collection of metadata, including: *'Title', 'Subject', 'Institution', 'Journal', 'Publisher', 'Country', 'Author', 'URLS', 'ArticleType', 'RetractionDate', 'RetractionDOI', 'RetractionPubMedID', 'OriginalPaperDate', 'OriginalPaperDOI', 'OriginalPaperPubMedID', 'RetractionNature', 'Reason', 'Paywalled'.*

3.1 Analysis of Retracted Dataset

Approximately 72% of the 8,087 retractions in our dataset originate from one of five countries (see Table 3.1). China contributed 39.7% of the total retractions, followed by the United States at 18%.

For a majority of articles, limited to no information about the reason for retraction is available in the dataset. In cases where that information is given, investigation by external

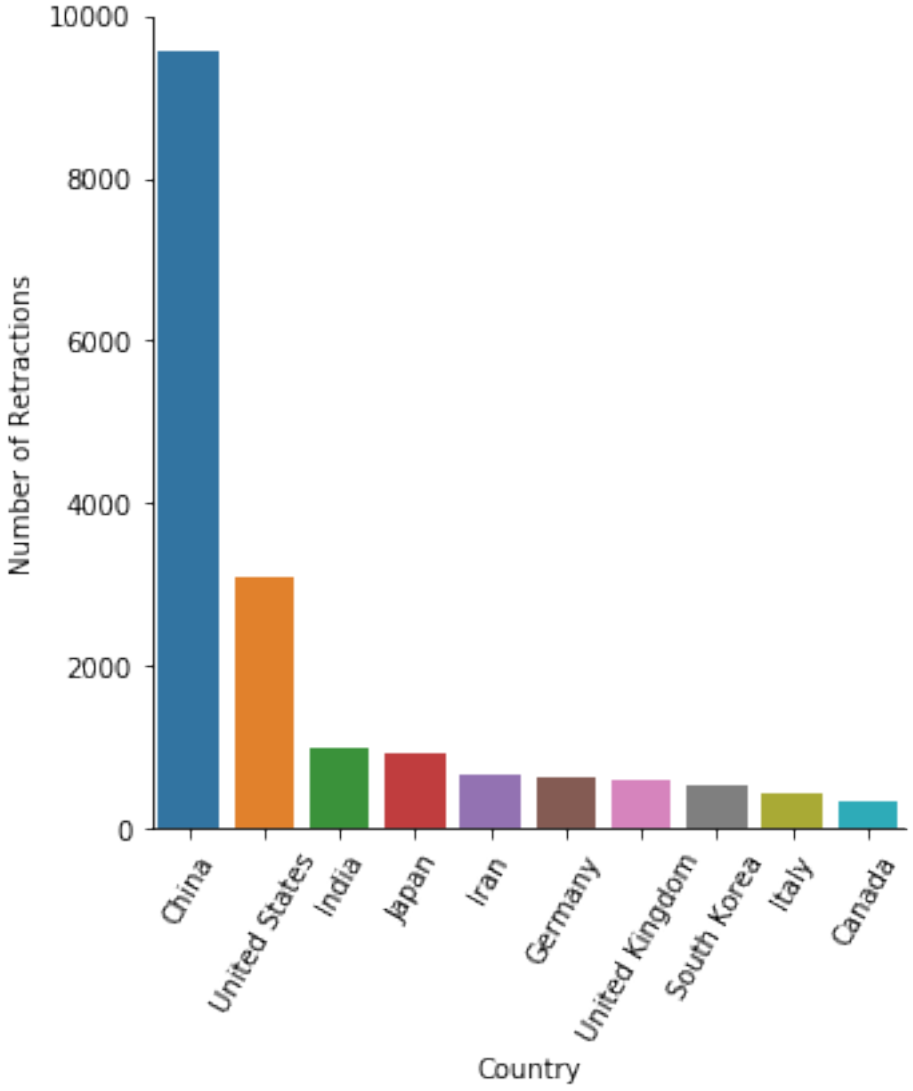

Figure 3.1: Number of retractions ordered by the top 10 countries.
Retractions by Country

parties such as journals, institutions, companies, etc., contribute to 27.7% of retractions. Malpractices such as plagiarism, duplication, falsification, fabrication, manipulation of data represent 37.3% (most malpractice is determined as the result of an investigation). Other prevalent reasons for retractions include breach of policy by authors, withdrawals by authors, and author misconduct (see Table 3.3). Of 27,471 authors appearing in the dataset, 500 contribute to 3,863 (of 8,087) retractions. Eighty-five authors have ten or more than ten retractions. This trend echoes similar findings reported in [36].

	China	United States	Japan	India	Germany
Count	3,211	1,462	460	392	314

Table 3.1: Top 5 number of retractions by country. Note that more than one country may be listed for a given record in the database.

Subject	Count
Health Sciences	5,952
Social Sciences	3,877
Humanities	376

Table 3.2: Number of retractions by subject. Note that more than one subject may be listed for a given record in the database.

The average time from date of publication to date of retraction in our dataset is 2 years. However, retraction time varies by subject. Average retraction time is 2.7 years for papers in HSC, as compared to 0.8 years in SOC and 1.7 years in HUM. We also observe a significant variation in the distribution of reasons for retractions across subjects. For example, retractions due to limited or no information contributed to 69% of retractions in SOC; the same reason contributed to only 14% of retractions in HSC. Similar observations were drawn in a study of retractions in the surgical literature [37].

3.2 Dataset for Classification

Of the 8,087 records, we further downsampled the records which have entries in PubMed. This choice to downsample to records available in PubMed is because abstracts and mesh terms available from PubMed can be used to search comparable negative samples. Of the records available in PubMed, we focus on records for which we can collect full-texts. Finally, we end up with 4,550 records of positive samples along with their full-texts for the classification task.

Reason for Retraction	Count
Limited or No Information	2,568
Investigation by Journal/Publisher	1,460
Investigation by Company/Institution	881
Duplication of Article	838
Withdrawal by author	673

Table 3.3: Top 5 reasons for retractions. Note that there may be more than one reason listed for a given record.

3.3 Negative Samples Collection

For classifier development and testing, a comparable set of non-retracted published articles (negative training samples) in a one-to-one mapping with retracted articles was collected such that:

- The negative sample was published within 3 years (before or after) the year of publication of the retracted sample.
- The negative sample most closely matches the retracted sample based on keywords (see below for details).

Keywords were retracted from papers using the TextRank algorithm proposed in [38]. TextRank uses a graph-based ranking model, which can be effectively used to extract keywords from text without the need for domain knowledge or annotated corpora. Extracted keywords were used to search for papers on similar topics around the same year of publication using the PubMed Entrez API[1]. The paper selected as the top match to each retracted paper, published within the three-year time window, was selected for inclusion in the negative training set. With collected negative samples and positive samples, our final dataset has $8,744$ records.

For both the records from Retraction Watch and the records selected from PubMed, we collected full-text PDFs. Full-text was used for developing features that are dependent of content of the published work such as use of p-values, sample sizes, etc. However,

[1]https://www.ncbi.nlm.nih.gov/home/develop/api/

these full-texts which are generally in PDF format needs to be preprocessed to be useful for developing features or classifiers

3.4 Preprocessing of full-texts

For both the records from Retraction Watch and the records selected from PubMed, we collected and preprocessed full-text PDFs. We experimented with several available conversion tools. While *pdftotext*[2] worked well for PDF to text conversion, it did not structure output in a usable way. Instead, to extract data from articles in a structured format, we used the GeneRation Of BIbliographic Data (Grobid) [39], which can segment PDF papers into TEI format, allowing programmatic access to various fields and sections of the paper. The GROBID output is further parsed using regular expression patterns (GROBID) and downstream feature extraction/development tasks.

[2]https://www.xpdfreader.com/pdftotext-man.html

Chapter 4

Features

We use a comprehensive set of features, including publication metadata and features derived from the full-text of published papers. Metadata features are pulled through public scholarly APIs. While, we make use of various mining tools including GROBID and *pdftotext* to extract pertinent information from full-text PDFs of published articles.

4.1 Metadata features

We leverage the Scopus[1], Crossref[2] and Semantic Scholar[3] datasets and tools to collect key measures related to the papers in our dataset.

Scopus[4] is one of the biggest metadata repositories that is publicly available. Furthermore, through its APIs, it provides metadata related metrics such as journal rankings. Crossref[5] structures metadata and represents relationships between metadata objects. Semantic Scholar[6] uses state of the art deep learning/machine learning methods to provide search tools for scholarly work. It also provides metrics and other semantics related extraction tools (e.g., results, intent). In this work, we aggregate multiple metadata

[1]https://www.scopus.com/
[2]https://www.crossref.org/
[3]https://www.semanticscholar.org/
[4]https://www.scopus.com/
[5]https://www.crossref.org/
[6]https://www.semanticscholar.org/

features from various scholarly APIs.

Pubmed provides a large collection of metadata from SBS/medicine related to scholarly work. Pubmed also provides annual and daily baseline dumps[7].

4.1.1 Lead author university rankings

GROBID output includes the author's first and last names and institutional affiliations. We use this information when available. When missing, we search for authors' affiliation information through Elsevier API. We then augment the first author's affiliation with an affiliation score, calculated using institutional rankings from Times Higher Education[8] as follows:

$$\text{Affiliation Score} = \begin{cases} 1 - \frac{\text{Rank}}{100} & \text{if Rank} < 100 \\ 0 & \text{otherwise} \end{cases}$$

We retain numeric ranks only for the top 100 universities and set the others to a default score.

4.1.2 Journal impact score

We use the SCImago Journal Rank (SJR) as the journal impact score, which is calculated as the average weighted citations per year divided by the total number of papers published in that journal over the past three years, where weight is determined by the prestige of the citing journal (see the SJR documentation for more details[9]).

$$SJR = \frac{\text{weighted average citations per year}}{\text{total documents published in three years}}$$

[7]https://www.nlm.nih.gov/databases/download/pubmed_medline.html
[8]https://www.timeshighereducation.com/world-university-rankings
[9]https://www.scimagojr.com/SCImagoJournalRank.pdf

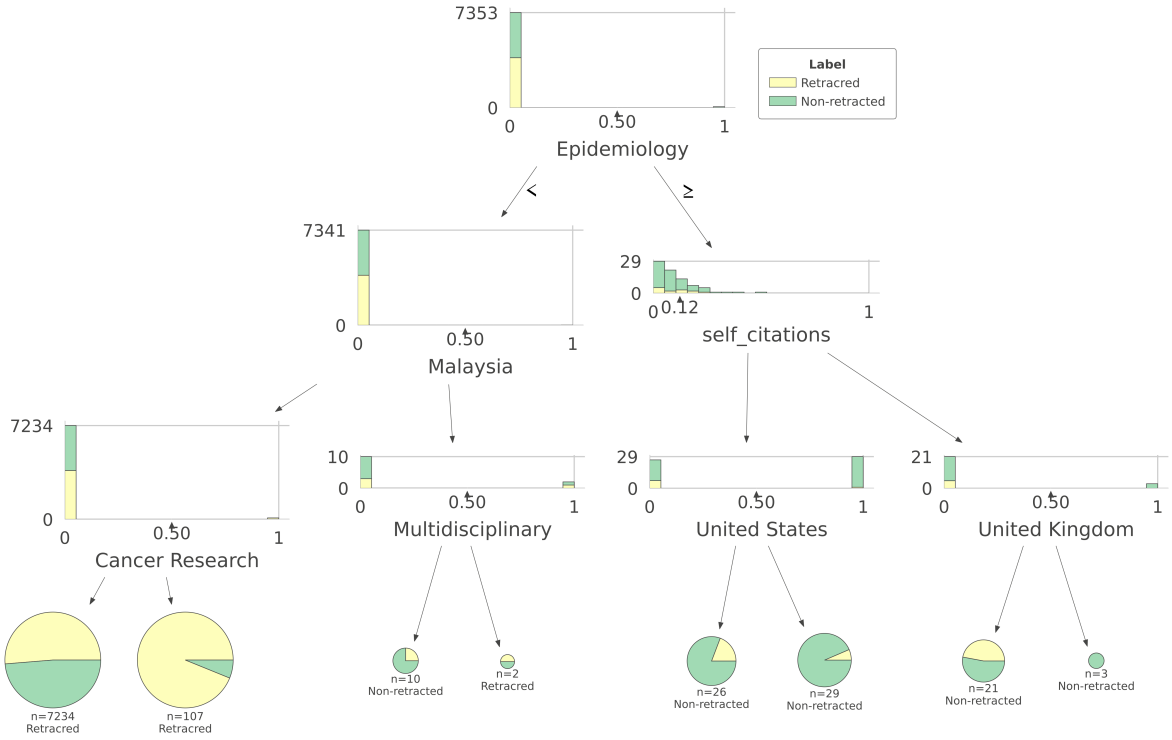

Figure 4.1: Decision Tree with depth=3, and country, subject, and self-citations as features.

4.1.3 Citation Count

We calculate citation count for a given paper as the total number of citations it has received. We collect citation counts from various sources and select the maximum count from that set. If a value is not available in any of these APIs, we set it to zero.

$$CitationCount = \max_{i=ScholarlyAPI} CitationCount_i$$

4.1.4 Citation Next

Citation Next [40] gives the average number of citations of a published work in the first three to five years after it has been published. It is calculated as:

$$CitationNext = \frac{\text{number of citations in the first n years}}{n}$$

4.1.5 Citation Velocity

The citation velocity represents the average rate at which a paper is cited in recent years, excluding self-citations [41]. The value is retrieved from the Semantic Scholar API. A detailed explanation of how this metric is calculated can be found in the Semantic Scholar documentation[10] .

4.1.6 Citation and Reference Intents

Semantic Scholar also provides the intent behind each citation and reference. A paper can be cited as background, methodology, results, etc. For a given paper, we count the number of citing papers of certain intent(s) by querying the paper's identifiers (title or DOI) against the Semantic Scholar API. Similarly, we count the number of references for each intent of the given paper and use them as features.

4.1.7 Open access

The open-access feature indicates whether the article can be accessed by any individual without a paywall. We collect this information from the Elsevier API and encode this flag as a binary feature.

4.1.8 Other Features

In addition to the features outlined above, we use other readily available standard metadata including: (i) subject area in which paper is published; (ii) country of the primary authors' affiliations; (iii) the number of references; (iv) number of authors; and, (v) title (we concatenate title along with abstract).

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

4.2 Full-Text Features

While metadata features give an overview of the paper, full-text features represent features that are much more content-specific. Specifically, we extract test statistics of experiments from full-text. These features are extracted using PDF conversion tools followed by various downstream feature extraction tasks.

4.2.1 p-values

p-values signifies the confidence level of a null hypothesis based on experiments. Full-texts of published work can be mined to extract p-values and various other test statistics. For this, we use *pdftotext* to extract textual information present in full-texts PDFs. Most of the papers in SBS fields follow standard formats to report p-values. For example, p-values are reported as $p < 0.01$, or $p = 0.1$, or $p > 0.5$, etc. We follow methods similar to [42] to extract p-values using various regex patterns.

Furthermore, we extract other features from the p-values identified using the regex patterns such as number of p-values, real-p: defined as the lowest p-values among all the extracted p-values, sign-$p \in \{>, <, =\}$: defined as the sign of the real-p, p-value range: defined as the difference between the highest and lowest p-values extracted from text. Some scholarly works publish p-values along with test statistics such as ANOVA, Chi-squared, etc. We use a binary feature that indicates whether the p-value is reported along with a test statistic is extended-p. For example, $F(200) = 13.8, p = 0.1$.

$$extended - p = \begin{cases} true & \text{if reported with test statistic} \\ false & \text{otherwise} \end{cases}$$

We use the the number of p-values with test statistic and the number of p-values without test statistics as features. In the future sections, we refer all the above p-value related features as p-value features rather than referring them individually.

4.2.2 Sample Size

Sample size is the number of observations made to determine the statistical significance of a hypothesis. Similar to p-value extraction, sample size can be extracted from a published article using regex patterns. In cases where test statistics are given, sample sizes can be calculated using various formulas based on the test statistic used. We use a combination of regex patterns and test statistic related formulas to extract sample sizes from a given paper.

4.2.3 Acknowledgements

The acknowledgment section of a published paper may contain funding information. We use AckExtract to extract named entities using state-of-the-art Named Entity Recognition techniques, followed by a relation-based entity classifier to determine if the work was funded by an organization [43].

4.2.4 Self Citations

Self-citation is common practice within the scientific community. Authors may cite their earlier works. The effects of self-citations and their significance for a paper's impact factor have been extensively studied [44, 45]. Authors publishing in high-impact journals have more self-citations when compared with authors usually publishing in lower-impact journals [46]. However, when self-citation ratios are considered, they observe high-impact journals have lower self-citation ratios when compared with lower-impact journals. We extract self-citations from the references section of full-text by matching author names and calculate the self-citation ratio. For matching, we used author names in the title section to compare with the author names in the references section using a fuzzy string matcher.

4.2.5 Abstract

The abstract section provides an overview of what the article is about and its area of study. Capturing the abstract information in a meaningful and effective way as a feature can play an important role in the classification task. In this work, we have experimented with various word embeddings to represent abstracts.

Doc2Vec Embeddings: Sentence embeddings learned via distributed representations are proven to be effective in sentence classification tasks [47]. Here, we experiment with these embeddings available as Doc2Vec in Gensim library [48].

BioSentVec embeddings: Along with Doc2Vec embeddings, we also experiment with BioSentVec embeddings proposed by [49]. BioSentVec is trained on large a large corpus of scholarly articles available in the PubMed database and clinical notes from MIMIC- III Clinical Database. The abstracts in our classification task are from a similar distribution on which BioSentVec is trained (Since all the records in our dataset are available in PubMed).

SciBERT embeddings: Bidirectional transformers have achieved state of the art results on most NLP tasks, including sentence classification. We experiment with sentence embeddings from SciBERT [50] embeddings obtained via bidirectional transformers trained on a large corpus of scholarly articles from Semantic Scholar. In our experiments, we use *[CLS]* token embeddings from SciBERT's output. In cases where abstract exceeded 512 tokens, we omitted the extra tokens for embeddings.

TFIDF: Term Frequency-Inverse Document Frequency (TFIDF) is a popular technique in information retrieval and machine learning. In our experiments, we use TFIDF of abstracts with removed stop words removed along with words stemmed. We also use TFIDF with reduced dimensions using TruncatedSVD [51].

Chapter 5
Classification Results

We formulate the task of retraction classification as follows: given access to a labeled set of training samples, $\{(x_i, y_i)\}_{i=1}^{n} \in \mathcal{X}_{train} \times \mathcal{Y}_{train}$, such that $x_i \in \mathbb{R}^n, y_i \in \{0, 1\}$ we aim to train a classifier $f : \mathcal{X} \to \mathcal{Y}$ with minimum classification error on unseen data i.e, $\mathcal{X}_{test} \times \mathcal{Y}_{test}$.

$$y_i = \begin{cases} 0 & \text{if retracted,} \\ 1 & \text{if non retracted} \end{cases}$$

We use random forest classifier [52] to support interpretability of results and good performance. All of our experiments were done using 100 trees as we didn't see much performance improvements over 100 trees. For experiments in Table 5.1, we used TF-IDF for representing abstracts. Note that we concatenate the title of the paper along with the abstract as a single feature. To further simplify the model for interpretability, we decompose the TFIDF matrix using randomized SVD [51] with 10 iterations to 15 dimensions. Randomized SVD is better suited for sparse matrices such as TFIDF. (We also experimented with PCA for dimensionality reduction, but dimensionality reduction using randomized SVD gave better results). For categorical variables in our dataset, i.e, *Subject, Country*, we use target encoding [53]. Target encoder takes into account the posterior probability of the target, given a categorical value and the prior probability of the target on the entire training set to encode categorical variables. We report 10-fold

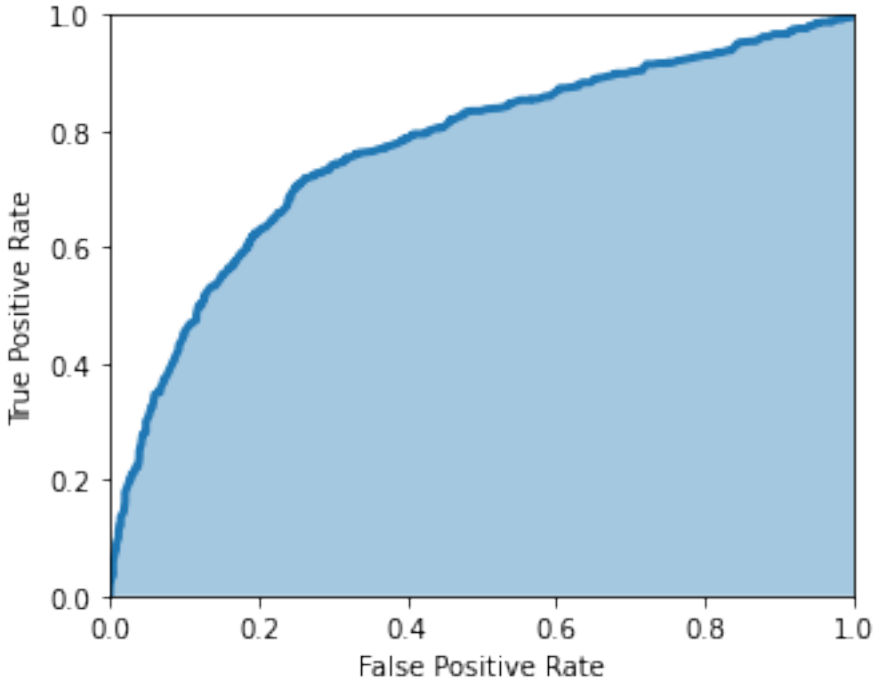

Figure 5.1: ROC Curve

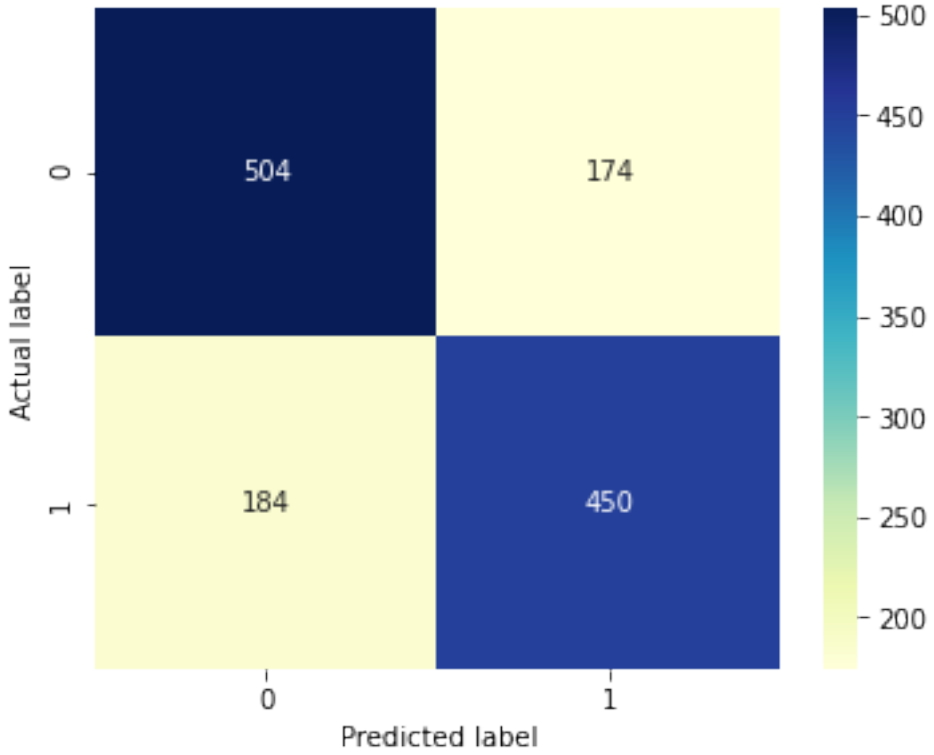

Figure 5.2: Confusion Matrix

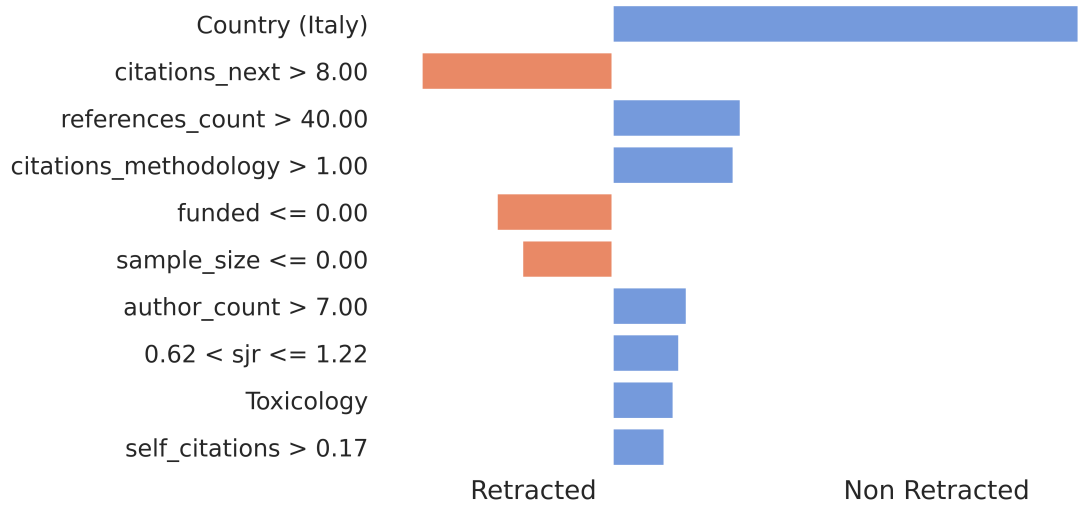

Figure 5.3: Plot showing features that contributed for Non-retraction classification of a sample

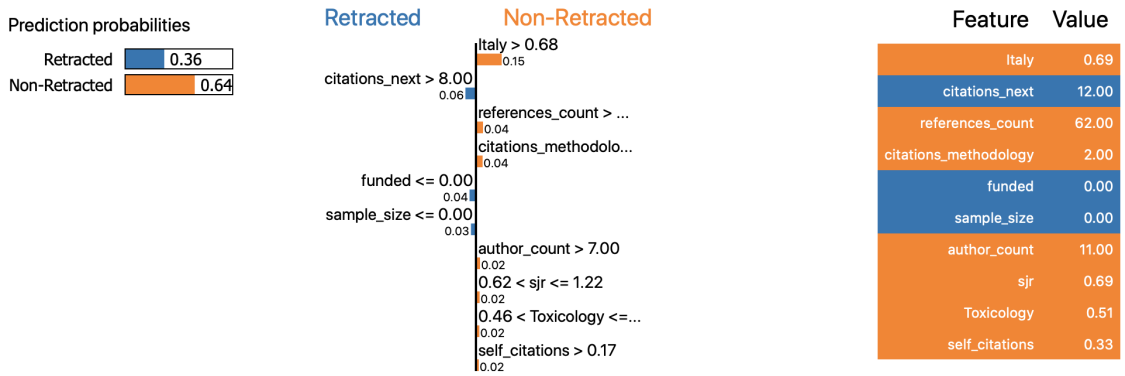

Figure 5.4: Plot showing features that contributed for retraction classification of a sample

cross-validation scores and scores on the train-test split (85% - 15%), see Table 5.1. For the train-test split, we report Area Under the Receiver Operating Characteristic (AUROC) of 78.1%. The ROC curve and a heat map of the confusion matrix are provided in Figure **??**.

A closer look at individual decision trees of our random forests reveals several interesting insights. For example, certain combinations of countries and subjects combined with other underlying feature distributions of such as low *SJR* and *University Rank* are more prone to retractions. On the other hand, certain combinations of countries and subjects with a high self-citation ratio are less likely to be retracted. This can be observed

in Figure 4.1. For the purpose of better visualization, we considered only three features: *Countries*, *Subjects* and *SJR*, and limited the depth of Decision Tree to 3. These three features together give an F1 score of 66%. Countries and Subjects are one-hot encoded for ease of understanding as opposed to target encoded for the scores in Table 5.1.

We further visualize a sample with actual and predicted label as non-retracted using Local Interpretable Model-Agnostic Explanations(LIME) [54] to present the effectiveness of our classifier. LIME explains an individual prediction by perturbing a sample and observing how the prediction changes around the given sample's perturbations. From Figure 5.3, we can observe that the non-retracted sample has more than seven authors with the primary author's affiliation, located in Italy. The paper has more than 40 references, which was cited more than once as methodology and a self-citation ratio greater than 0.17. The SJR score of the journal where the paper is published falls in the interval $[0.62, 1.2]$. All these attributes contributed towards non-retracted classification confidence. While the overall prediction is non-retracted, having no funding agency acknowledged, no sample size information, and *citation_next* value greater than eight are seen as attributes that could lead to retraction. Note that this visual analysis is particular to a sample and does not represent the global feature importance, and is meant for a high-level intuition of how various features can meaningfully impact a published work's confidence.

	10-Fold Cross Valid.	**Train-test Split**
Accuracy	73.65	73.32
Precision	74.32	71.54
Recall	68.70	72.00
F1	71.37	71.77

Table 5.1: Random Forest Classifier performance for Accuracy, Precision, and Recall scores, averaged for 10-fold cross validation and train-test split

5.1 Ablation Studies

We completed an ablation study to identify features (or combinations of features) that are instrumental in identifying retracted papers. Table 5.2 shows the result of this investigation. Metadata features alone give an F1 score of 67%, while full-text features alone result in an F1 score of 63%. Combined together, metadata and full-text features help improve performance to an F1 score of 71%. The importance of full-text features can also be observed by excluding abstract, self-citations, and p-value features individually. Excluding abstract, self-citations ratio, and p-value separately doesn't lead to a significant drop in F1-score, but together they drop the F1-score to 67.7%.

We examine the importance of each feature by excluding each from the overall features and also measuring the performance of each feature individually. In Table 5.2, the country of the primary author has the most predictive power. Excluding the country from the overall feature list hurts the F1-score significantly. Individually, *SJR, abstract, country* give the best performance out of all metadata features. Similarly, the TFIDF of the abstracts gives the best performance of all the full-text features. We reduced the dimension of the TFIDF vector from 34,000 to 15 using Truncated SVD without a significant drop in performance. The best score is achieved by using all the features.

In regards to individual features, from Table 5.2 we note that features such as *self-citation* alone cannot achieve any separability. However, when combined with other features, they provide predictive power to the classifier Figure 4.1. University rank individually provides almost no separability. The university rank of $8,535$ records is set to default value 0; this suggests exploring better methods to encode affiliation information. $3,130$ records in our dataset have open access (open access flag set to 1), this feature exhibits almost zero correlation(-0.017) with *retracted vs. non-retracted* label. This suggests that open access of published articles is not an indicator of a scholarly work's confidence.

	Accuracy	Precision	Recall	F1
Individual Features:				
Abstract$_{TF-IDF}$	67.14	64.71	69.01	66.76
Abstract$_{SciBERT}$	65.69	64.48	63.06	63.75
Abstract$_{SVD_{n=15}}$	65.05	63.51	63.30	63.39
Country	65.93	66.02	59.60	62.57
Abstract$_{BioSentVec}$	65.24	64.76	60.26	62.40
SJR	66.22	69.37	52.68	59.85
Subject	63.53	64.43	53.58	58.39
Cite. Next	57.07	56.02	48.59	51.95
Cite. Background	56.06	54.70	48.29	51.25
Cite. Results	56.06	54.70	48.29	51.25
Author Count	51.66	49.57	50.09	49.44
p**-value features**	53.61	51.84	44.09	47.62
Self-Cite.	53.29	51.62	38.57	44.14
Ref. Background	54.94	54.42	36.45	43.53
Abstract$_{DOC2VEC}$	50.69	48.12	37.18	41.89
Funded	54.87	54.58	34.00	41.87
Ref. Methodology	54.56	54.83	29.14	37.97
Cite. Methodology	54.09	54.14	25.87	34.98
Ref. Results	52.72	51.61	20.02	28.62
Uni. Rank	52.41	59.39	1.64	3.20
Open Access	52.14	0.00	0.00	0.00
Particular Feature Excluded:				
Cite. Next	73.42	74.25	68.08	71.00
Uni. Rank	73.36	74.22	67.99	70.94
Open Access	73.30	74.23	67.78	70.83
p**-Value features**	73.28	74.37	67.49	70.73
Author Cnt.	73.18	74.18	67.36	70.59
Self-Cite.	73.07	73.97	67.56	70.58
Abstract	72.72	72.95	68.37	70.58
Funded	73.01	74.07	67.13	70.41
Subject	72.64	73.55	66.97	70.07
Country	71.01	70.82	67.10	68.87
Overall Features:				
Metadata	71.73	74.74	61.96	67.70
Full-text	65.31	63.74	63.78	63.72
All Features	73.65	74.32	68.70	71.37

Table 5.2: Ablation study results. Ordered by individual feature performance, performance with particular feature excluded from all the features and overall performance results.

Chapter 6
Discussion and Conclusion

6.1 Discussion

One of the important questions that arise when building any particular model is whether increasing the data improves the classification accuracy. This can be checked by down-sampling the data and comparing the metrics of the down-sampled class with that of the original dataset. Another interesting experiment would be to check if individual subjects (SOC, HUM, HSC) classification accuracies vary across subjects. Since we have observed differences in data distribution of different subjects, it would be interesting to verify if there are any patterns in a particular subject that help in better/worse prediction of that subject when compared to others.

Another important observation is how the country distribution in retracted papers is skewed towards China. This provides classifier predictive bias that the model can leverage. It is important to note that China also produces a number of high-quality papers. In order to test the robustness of our classifier and features, we can collect a different negative sample set with a country distribution similar to that of the retracted set.

Another interesting observation is how our task is similar to classifying reproducibility of a published work [55]. It can be an interesting experiment to check if the dataset

from the reproducibility project can be used to pre-train our classifier or vice versa. Alternatively, we can combine the datasets to formulate the problem as classifying the quality of a published paper. Or a more challenging formulation is a multi-task classification problem where we predict the paper as either reproducible, retracted or neither.

6.2 Conclusion

In this work, we present initial evidence for the utility of supervised approaches for the assessment of retracted scholarly work. Using metadata as well as features derived from the full-text for a subset of retracted papers in the social and behavioral sciences, we develop a random forest classifier to predict retraction in new samples. Looking ahead, we might assume that signals of credibility and concern will vary across scientific domains. And that further studies in ML-enabled understanding of retraction will, therefore, likely need to be undertaken by interdisciplinary teams. We suggest that yet more sophisticated features capturing argument structure, experimental conditions, and corroborations across the literature will be important steps for work in this direction.

Bibliography

[1] Collaboration, O. S. et al. (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

[2] Camerer, C. F., A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer, et al. (2018) "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015," *Nature Human Behaviour*, **2**(9), pp. 637–644.

[3] Klein, R. A., M. Vianello, F. Hasselman, B. G. Adams, R. B. Adams Jr, S. Alper, M. Aveyard, J. R. Axt, M. T. Babalola, Š. Bahník, et al. (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.

[4] Steen, R. G., A. Casadevall, and F. C. Fang (2013) "Why has the number of scientific retractions increased?" *PloS one*, **8**(7), p. e68397.

[5] Fanelli, D. (2013) "Why Growing Retractions Are (Mostly) a Good Sign," *PLoS Medicine*, **10**(12), p. e1001563.
URL https://dx.plos.org/10.1371/journal.pmed.1001563

[6] Brainard, J. (2018) "Rethinking retractions," *Science*, **362**(6413), pp. 390–393, https://science.sciencemag.org/content/362/6413/390.full.pdf.
URL https://science.sciencemag.org/content/362/6413/390

[7] Casadevall, A., R. G. Steen, and F. C. Fang (2014) "Sources of error in the retracted scientific literature," *The FASEB Journal*, **28**(9), pp. 3847–3855.

[8] Hesselmann, F., V. Graf, M. Schmidt, and M. Reinhart (2017) "The visibility of scientific misconduct: A review of the literature on retracted journal articles," *Current sociology*, **65**(6), pp. 814–845.

[9] Bornmann, L. and R. Mutz (2015) "Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references," *Journal of the Association for Information Science and Technology*, **66**(11), pp. 2215–2222, https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23329.
URL https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23329

[10] Garfield, E. et al. (1994) "The impact factor," *Current contents*, **25**(20), pp. 3–7.

[11] Seglen, P. O. (1997) "Why the impact factor of journals should not be used for evaluating research," *Bmj*, **314**(7079), p. 497.

[12] Bordons, M., M. T. Fernández, and I. Gómez (2002) "Advantages and limitations in the use of impact factor measures for the assessment of research performance," *Scientometrics*, **53**(2), pp. 195–206.
URL https://doi.org/10.1023/A:1014800407876

[13] Stodden, V., M. McNutt, D. H. Bailey, E. Deelman, Y. Gil, B. Hanson, M. A. Heroux, J. P. Ioannidis, and M. Taufer (2016) "Enhancing reproducibility for computational methods," *Science*, **354**(6317), pp. 1240–1241.

[14] Perkel, J. M. (2018) "A toolkit for data transparency takes shape," *Nature*, **560**(7718), pp. 513–516.

[15] Oransky, I. and A. Marcus (2012), "Retraction watch," .
URL http://retractiondatabase.org/RetractionSearch.aspx?

[16] Foster, E. D. and A. Deardorff (2017) "Open science framework (OSF)," *Journal of the Medical Library Association: JMLA*, **105**(2), p. 203.

[17] Galak, J., R. A. LeBoeuf, L. D. Nelson, and J. P. Simmons (2012) "Correcting the past: Failures to replicate psi." *Journal of personality and social psychology*, **103**(6), p. 933.

[18] Ritchie, S. J., R. Wiseman, and C. C. French (2012) "Failing the Future: Three Unsuccessful Attempts to Replicate Bem's 'Retroactive Facilitation of Recall' Effect," *PLOS ONE*, **7**(3), pp. 1–5.
URL https://doi.org/10.1371/journal.pone.0033423

[19] Kerwin, J., I. Khan, et al. (2020) "Replication Study: A coding-independent function of gene and pseudogene mRNAs regulates tumour biology," *eLife*, **9**, p. e51019.

[20] Repass, J. et al. (2018) "Replication Study: Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma," *Elife*, **7**, p. e25801.

[21] Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, et al. (2016) "Evaluating replicability of laboratory experiments in economics," *Science*, **351**(6280), pp. 1433–1436.

[22] Bennett, C., L. M. Chambers, L. Al-Hafez, C. M. Michener, T. Falcone, M. Yao, and V. Berghella (2020) "Retracted articles in the obstetrics literature: lessons from the past to change the future," *American Journal of Obstetrics & Gynecology MFM*, **2**(4), p. 100201.
URL http://www.sciencedirect.com/science/article/pii/S2589933320301701

[23] Coudert, F.-X. (2019) "Correcting the Scientific Record: Retraction Practices in Chemistry and Materials Science," *Chemistry of Materials*, **31**, pp. 3593–3598.

[24] Dal-Ré, R. (2019) "Analysis of retracted articles on medicines administered to humans," *British Journal of Clinical Pharmacology*, **85**(9), pp. 2179–2181, https://bpspubs.onlinelibrary.wiley.com/doi/pdf/10.1111/bcp.14021.
URL https://bpspubs.onlinelibrary.wiley.com/doi/abs/10.1111/bcp.14021

[25] Nogueira, T. E., A. S. Gonçalves, C. R. Leles, A. C. Batista, and L. R. Costa (2017) "A survey of retracted articles in dentistry," *BMC Research Notes*, **10**(1), p. 253.
URL https://doi.org/10.1186/s13104-017-2576-y

[26] Pantziarka, P. and L. Meheus (2019) "Journal retractions in oncology: a bibliometric study," *Future Oncology*, **15**(31), pp. 3597–3608, pMID: 31659916, https://doi.org/10.2217/fon-2019-0233.
URL https://doi.org/10.2217/fon-2019-0233

[27] Mistry, V., A. Grey, and M. J. Bolland (2019) "Publication rates after the first retraction for biomedical researchers with multiple retracted publications," *Accountability in Research*, **26**(5), pp. 277–287, pMID: 31025884, https://doi.org/10.1080/08989621.2019.1612244.
URL https://doi.org/10.1080/08989621.2019.1612244

[28] Mott, A., C. Fairhurst, and D. Torgerson (2019) "Assessing the impact of retraction on the citation of randomized controlled trial reports: an interrupted time-series analysis," *Journal of Health Services Research & Policy*, **24**(1), pp. 44–51, pMID: 30249142, https://doi.org/10.1177/1355819618797965.
URL https://doi.org/10.1177/1355819618797965

[29] Suelzer, E. M., J. Deal, K. L. Hanus, B. Ruggeri, R. Sieracki, and E. Witkowski (2019) "Assessment of Citations of the Retracted Article by Wakefield et al With Fraudulent Claims of an Association Between Vaccination and Autism," *JAMA Network Open*, **2**(11), pp. e1915552–e1915552, https://jamanetwork.com/journals/jamanetworkopen/articlepdf/2755486/suelzer_2019_oi_190588.pdf.
URL https://doi.org/10.1001/jamanetworkopen.2019.15552

[30] Chan, M., C. Jones, and D. Albarracín (2017) *Countering false beliefs: An analysis of the evidence and recommendations of best practices for the retraction and correction of scientific misinformation*, Oxford University Press, pp. 341–350.

[31] Dal-Ré, R., L. M. Bouter, D. Moher, and A. Marušić (2020) "Mandatory disclosure of financial interests of journals and editors," *BMJ*, **370**, https://www.bmj.com/content/370/bmj.m2872.full.pdf.
URL https://www.bmj.com/content/370/bmj.m2872

[32] Horton, J., D. Krishna Kumar, and A. Wood (2020) "Detecting academic fraud using Benford law: The case of Professor James Hunton," *Research Policy*,

49(8), p. 104084.
URL http://www.sciencedirect.com/science/article/pii/S0048733320301621

[33] Lesk, M., J. B. Mattern, and H. M. Sandy (2019) "Are papers with open data more credible? An analysis of open data availability in retracted PLoS articles," in *International Conference on Information*, Springer, pp. 154–161.

[34] Dinis-Oliveira, R. J. (2020) "COVID-19 research: pandemic versus "paperdemic", integrity, values and risks of the "speed science"," *Forensic Sciences Research*, **5**(2), pp. 174–187, https://doi.org/10.1080/20961790.2020.1767754.
URL https://doi.org/10.1080/20961790.2020.1767754

[35] Soltani, P. and R. Patini (2020) "Retracted COVID-19 articles: a side-effect of the hot race to publication," *Scientometrics*, **125**(1), pp. 819–822.
URL https://doi.org/10.1007/s11192-020-03661-9

[36] Brainard, J. and J. You (2018) "What a massive database of retracted papers reveals about science publishing's 'death penalty'," *Science*, **25**(1), pp. 1–5.

[37] King, E. G., I. Oransky, T. E. Sachs, A. Farber, D. B. Flynn, A. Abritis, J. A. Kalish, and J. J. Siracuse (2018) "Analysis of retracted articles in the surgical literature," *The American Journal of Surgery*, **216**(5), pp. 851–855.

[38] Mihalcea, R. and P. Tarau (2004) "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 404–411.

[39] Lopez, P. (2009) "GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications," ECDL'09, Springer-Verlag, Berlin, Heidelberg, pp. 473–474.
URL http://dl.acm.org/citation.cfm?id=1812799.1812875

[40] 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.

[41] 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

[42] Nuijten, M. B., C. H. Hartgerink, M. A. van Assen, S. Epskamp, and J. M. Wicherts (2016) "The prevalence of statistical reporting errors in psychology (1985–2013)," *Behavior research methods*, **48**(4), pp. 1205–1226.

[43] Wu, J., P. Wang, X. Wei, S. Rajtmajer, C. L. Giles, and C. Griffin (2020) "Acknowledgement Entity Recognition in CORD-19 Papers," in *Proceedings of the First Workshop on Scholarly Document Processing*, Association for Computational Linguistics, Online, pp. 10–19.
URL https://www.aclweb.org/anthology/2020.sdp-1.3

[44] Renata, T. (1977) "SELF-CITATIONS IN SCIENTIFIC LITERATURE," *Journal of Documentation*, **33**(4), pp. 251–265.
URL https://doi.org/10.1108/eb026644

[45] Wolfgang, G., T. Bart, and S. Balázs (2004) "A bibliometric approach to the role of author self-citations in scientific communication," *Scientometrics Scientometrics*, **59**(1), pp. 63–77.
URL https://akjournals.com/view/journals/11192/59/1/article-p63.xml

[46] Anseel, F., W. Duyck, W. De Baene, and M. Brysbaert (2004) "Journal Impact Factors and Self-Citations: Implications for Psychology Journals." **59**(1), pp. 49–51.

[47] Le, Q. V. and T. Mikolov (2014) "Distributed Representations of Sentences and Documents," in *International Conference on Machine Learning*.

[48] Řehůřek, R. and P. Sojka (2010) "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, Valletta, Malta, pp. 45–50, http://is.muni.cz/publication/884893/en.

[49] Chen, Q., Y. Peng, and Z. Lu (2019) "BioSentVec: creating sentence embeddings for biomedical texts," *2019 IEEE International Conference on Healthcare Informatics (ICHI)*.
URL http://dx.doi.org/10.1109/ICHI.2019.8904728

[50] Beltagy, I., K. Lo, and A. Cohan (2019) "SciBERT: A Pretrained Language Model for Scientific Text," Association for Computational Linguistics, Hong Kong, China, pp. 3615–3620.
URL https://www.aclweb.org/anthology/D19-1371

[51] Halko, N., P. G. Martinsson, and J. A. Tropp (2011) "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions," *SIAM Rev.*, **53**(2), p. 217–288.
URL https://doi.org/10.1137/090771806

[52] Breiman, L. (2001) "Random forests," *Machine Learning*, **45**(1), pp. 5–32.
URL https://doi.org/10.1023/A:1010933404324

[53] Micci-Barreca, D. (2001) "A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems," *ACM SIGKDD Explorations Newsletter*, **3**(1), pp. 27–32.

[54] Ribeiro, M. T., S. Singh, and C. Guestrin (2016) ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144.

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