Predicting retracted research

Aaron HA Fletcher[a], Mark Stevenson[a]

[a]*Department of computer science, The University of Sheffield, Sheffield, S1 4DP, United Kingdom*

Abstract

Retracting published research is an important safeguard against the dissemination of flawed or fraudulent scientific information. However, detecting problematic research prior to publication remains a challenge. This paper describes the creation of a novel dataset and machine learning models to predict retracted articles. The data set combines information from the Retraction Watch database and the OpenAlex API, including article metadata, abstracts, and citation metrics. A total of 16,224 articles (8,112 retracted and 8,112 nonretracted) published between 2000-2020 were included. Several machine learning models were trained on this data, with a gradient boosting approach achieving the best precision (0.691). An ablation study revealed that the abstract of the article was the most important feature for classification for the accuracy, recall, and F1 score metric. First Author Countries was the more important feature for feature-based classifers with the Precision Metric. This work demonstrates the potential for using machine learning to assist in identifying problematic research during the peer review process, though further improvements in model performance are needed before practical application. The data set and code are made publicly available to support future work in this area.

Keywords: Retraction prediction, Machine Learning, Scientific publishing

1. Introduction

Retracting journal articles is a crucial safeguard against disseminating inaccurate or unreliable information. Conversely, the number of journal retrac-

Email addresses: ahafletcher1@sheffield.ac.uk (Aaron HA Fletcher), mark.stevenson@sheffield.ac.uk (Mark Stevenson)

tions can act as a proxy for failures within the publishing process, indicating instances where previous safeguards, such as peer review and editorial oversight, have failed to prevent the publication of flawed research. Numerous studies have successfully and repeatedly demonstrated an increasing trend in journal retractions [1, 2]. However, given the constraints of the current peer review system and the growing ability of natural language generation tools, preventing the publication of inaccurate or unreliable information versus providing a venue for the dissemination of research remains a challenging balance [3].

Abstractly, the decision to publish research is a binary classification task. The reviewer(s) act as a function that classifies an input (research) into two classes: to accept or not accept. This process closely resembles a text classification task in natural language processing (NLP), where the text is categorised into classes based on a function (such as sentiment) [4, 5]. Recent advances in NLP, including word embeddings, deep neural networks, and transformer architectures, have demonstrated considerable success in text classification across various domains. However, all supervised machine learning approaches fundamentally rely on labelled datasets, which to date have not been available.

Seeking the automatisation of identification of flawed research is important because of the potential benefits it could bring: warning the peer reviewer of any potential retraction risk before deciding to publish or precluding flawed research circulation. Although retracted articles are not void of academic usefulness, as they can be used to dismiss prior domain knowledge or direct future research areas, the valid use of retracted articles hinges on whether the end user is aware of an article's retraction status, which given the differing approach on how works are retracted, is not always clear. Continued improper research publication can have severe consequences, not just for the authors but also for the journal's reputation and the domain's integrity.

1.1. Existing Literature

Retractions occur for various reasons, broadly categorised into two main groups: (1) honest errors in otherwise ethically conducted research (estimated 73.5% of PubMed retractions between 2000 and 2010), and (2) improper or fraudulent research practices, including data fabrication, plagiarism, or false authorship claims (estimated 26.6% of PubMed retractions in the same period) [2]. However, these proportions can vary significantly between disciplines, as evidenced by the prevalence of misconduct-related retractions in BioMed Central journals [6]. This conflicting evidence likely stems from the challenges in accurately determining researchers' motivations, resulting in imprecise classification criteria. From the perspective of researcher end-users, the specific reasons for retraction may be less critical than the fact that retracted research is inherently unreliable, following the principle "garbage in, garbage out" [7]. Moreover, retraction reasons are determined retrospectively and would not be available when initially classifying a publication's risk of retraction.

Limited research exists on the demographics of authors producing retracted articles. Studies have found a significant association between first authors from lower-income countries and retractions due to plagiarism, suggesting global variations in retraction reasons [8]. Interestingly, contrary to the global trend, fraudulent research is more prevalent in the United States, with over half (53%) of fraudulent articles authored by "repeat offenders"[1]. These findings suggest that demographic characteristics, such as an author's name or country of origin, could be valuable features in a retraction classification dataset.

A 2022 study that examined retracted medical articles using the Retraction Watch dataset, Web of Science, and journal citation reports highlighted the growing phenomenon of paper mill retractions. The study reported an increase in retractions related to paper mills, predominantly associated with China [9]. The median time to retraction was two years, decreasing as the impact factor of the journal increased. Another study in 2021, also focused on the medical domain, revealed differences in the types of articles retracted, with 83.8% being original research and 8.6% being "meta" research [10]. These findings suggest that characteristics such as the impact factor of a journal, the country of origin, and the type of article may contain valuable information for modelling retractions.

To date, no research has investigated the plausibility of machine learning modelling in the prediction of these retracted articles.

1.2. Paper Contributions

- A novel open-source data set that can be used to model the prediction of retracted articles.

- The creation of classifier models to classify if an article is retracted.

2. Dataset Generation

Retraction watch is a human-validated retraction dataset. Retraction watch is compiled from various sources, including journal databases, institutional reports, social media, and direct tips [11, 12]. Although not exhaustive due to stealth "retractions" [13], it provides partial metadata for some retracted articles, such as title, journal, publisher, and author. OpenAlex, an open online catalogue of works, similar to Scopus and Web of Science [14], aggregates data from multiple sources monthly. The combination of these data sets was used to create a single data set suitable for predicting article retractions.

2.1. Dataset Generation

The Retraction Watch dataset was obtained and queries to the OpenAlex API occurred on 24/07/2024. All data retrieved are available in the GitHub repository data folder[1].

2.1.1. Inclusion/Exclusion Criteria

Unique journals in the retraction watch dataset were calculated after applying the journal and work exclusion criteria listed in Tables 3 and 4. Retractions were limited to a 20 year period from 2000 to 2020 due to the median lag of the works being retracted being 2 years, the lack of retracted works before this date and the increased use of natural language technologies subsequent to this period; see Figures A.6 and A.6. This generated a list of journals with retractions and retracted articles.For each journal, title, works count, citation count and H index features were recorded.

Only articles and review works were used within the dataset, with the type being determined by OpenAlex's *"type"* field. Conference papers were excluded due to the previously reported mass retraction of conference papers undertaken by the Institute of Electrical and Electronic Engineers between 2009 and 2011 (having pulled over 10,000 such papers in the past two decades) [15]. For each retracted article, another article was sampled randomly from the year of the retracted article also did not meet the works exclusion criteria outlined in Table 4, was not included in the retraction watch dataset and who's OpenAlex API flag of *'is_retracted'* was False. Unretracted

[1] `https://github.com/afletcher53/RetractionWatch`

works's who's title contained the keywords *"retraction", "retraction:", "withdrawn", "correction", "erratum", "retracted", "withdrawal", "conclusion", "editorial", "contributions", "commentary", "contributors"* were not eligible for sampling. From all works (retracted and unretracted), any were dropped if the abstract or title contained the words "elsevier", "notice", "editor", "editors", "publisher". For further information on works exclusion criteria, see Table 4.

For each work (retracted and non-retracted), the following features were recorded from OpenAlex:

1. Abstract Inverted Index
2. Publication Date
3. Primary Topic
4. First Author
5. Institution
6. Citation Count
7. First Author Countries
8. Is Retracted Flag
9. Article Type

2.1.2. Preprocessing

All textual features were preprocessed by converting them to lowercase and eliminating non-ASCII characters, HTML tags, numbers, additional symbols, and whitespace. The words listed in Table 2 were removed from the textual characteristics. Each data point was labelled as 0 (unretracted) or 1 (retracted). Finally, all features were concatenated with a descriptor preceding the values, resulting in the data format shown in Table 1. To balance the data set, the nonretracted works were randomly undersampled to match the number of retracted works. The data set (n=16224) was randomly assigned to three groups: test (20%, n=3245), train (64%, n=10383), and validation (16%, n=2596). For all feature-based classification models, all inputs were vectorised using a count vectorizer (max ngrams = 1). The resulting count vectors were then adjusted using saturated term frequency-inverse document frequency (B = 0.3 and K1 = 2). For contextual language understanding models (i.e., BERT), all inputs were tokenised using the *"bert-based-uncased"* model, with a maximum token length of 512 [16].

Table 1: Examples of generated data.

Label	**Input String**
1	**title** the feasibility of improving impact resistance and strength properties of sustainable concrete composites by adding waste metalized plastic fibres first **first author** hossein mohammadhosseini **first author countries** MY **primary topic** fiber reinforced concrete in civil engineering abstract **abstract** waste plastic results in waste discarding disaster and consequently cause significant harms to the environment the utilisation of industrial wastes production sustainable concrete has attracted much consideration recent years because lowcost materials along with saving a place for landfill purposes also enhance performance concrete in this paper feasibility metalized wmp fibres palm oil fuel ash pofa composites was investigated by assessing impact resistance strength properties six mixes containing wmp varying from length mm were made ordinary portland cement opc a different six mixtures same fibre content made where pofa substituted [Abstract truncated for brevity] **citated by count** 82 **publication date** 2018-04-01
0	**title** synergetic photoluminescence enhancement of monolayer mosvia surface plasmon resonance and defect repair **first author** yi zeng **first author countries** CN **primary topic** twodimensional materials **abstract** A the weak lightabsorption and low quantum yield qy in monolayer mos are great challenges for the applications of this material practical optoelectronic devices here we report on a synergistic strategy to obtain highly enhanced photoluminescence pl by simultaneously improving intensity electromagnetic field around qy mos selfassembled submonolayer au nanops underneath bistrifluoromethanesulfonimide tfsi treatment surface used boost excitation qy respectively an enhancement factor pl as high is achieved mechanisms analyzed inspecting contribution spectra from a excitons a trions under different conditions our study takes further step developing highperformance devices based **citated by count** 10 **publication date** 2018-01-01

Table 2: Banned Words / Phrases removed from corpus.

retraction	retracted	retract
retractionwatch	retraction watch	removed
withdrawn	withdrawal	withdraw
retracted article	article	

Table 3: Journal Exclusion Criteria.

Criteria	**Description**
CrossRef	If journal was not included in CrossRef's journal title list.
Works Count	If work count (*based on OpenAlex API*) - total retraction count (*based on Retraction Watch Dataset*) < Sample Size (*1*)
Retraction Count	If journal total retractions < 5 (*determined by the Retraction Watch dataset*).

Table 4: Work Exclusion Criteria.

Criteria	Description
Retracted Works	If Retraction Watch parameter *'ArticleType'* not in {Research Article, Conference Abstract/Paper, Clinical Study, Review Article, Case Report, Meta-Analysis}
Retracted Works/Non-retracted works	If OpenAlex *'source'* not in {Conference, Journal} and *'type'* not in {article, review}
English Language	Work excluded if OpenAlex API *'language'* value not 'en'
ISSN Data	If OpenAlex API *'issn'* value not available
OpenAlex ID	If OpenAlex API *'id'* value not available
Article Type	If OpenAlex API *'type'* value not in {article, review, conference-paper}
Publication Year	If *'publication_year'* OpenAlex API value not available
Publication Year Minimum	If *'publication_year'* OpenAlex API value < 2000
Publication Year Maximum	If *'publication_year'* OpenAlex API value > 2020
Reformulated Abstract Length	If < 5 words

3. Investigations

3.1. Dataset overview and characteristics

The generated data set maintained the trend of increasing retraction counts per year observed in the original Retraction Watch dataset, as illustrated in A.5, A.8 & A.8. When normalised between 0-1, the root mean square error between the generated and original data sets was 0.106, with the similarity shown in Figure 1. Shockingly, within this data set, 7.54% of the articles marked as retracted by the Retraction Watch data set were not marked as retracted by OpenAlex.

For all statistical tests $\alpha = 0.05$. Analysis of correlations between journal features revealed two notable findings:

1. A weak, significant positive correlation between the work count log and the retraction count log (Pearson correlation coefficient 0.065, p-value < 0.05), as shown in Figure A.10.

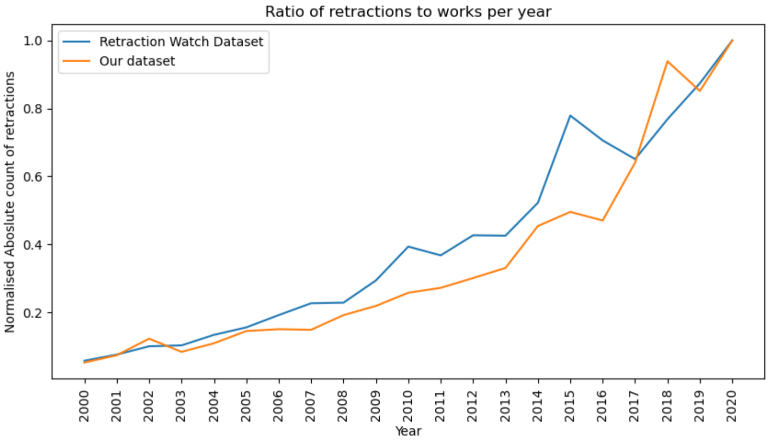

Figure 1: Comparison of our dataset and Retraction Watch Dataset when normalised between 0 and 1

2. A strong negative correlation between the retraction count and log of h-index (Pearson correlation coefficient -0.656, p-value < 0.05), as illustrated in Figure A.11.

3.2. Model Evaluation for Retraction Prediction

Various machine learning models were evaluated to determine if prediction of retraction was possible. The generated data set was used to train the models in the training data set and evaluate them with the test data set. Default settings from the Sklearn package or Huggingface were used unless specified, and random states were set to 42 [17].

The following models were used:

- Multi-layer Perceptron classifier (max_iter = 1000)
- Logistic Regression (max_iter = 1000)
- Decision Tree
- Random Forest
- Support Vector Machine (SVM) (kernel = rbf)
- Gradient Boosting
- XGBoost (n_estimators = 100, learning_rate = 0.1)

- AdaBoost
- BERT (Pretrained model = bert-base-uncased, AdamW optimizer with learning rate 2e-5, fine-tuned for 5 epochs on the training data)
- Llama 3.1 (Pretrained model = unsloth/llama-3-8b-bnb-4bit, AdamW optimizer with learning rate 1e-4, fine-tuned for 2 epoch on training data with early stopping).
- Gemma 2 (Pretrained model = unsloth/gemma-2-9b, AdamW optimizer with learning rate 1e-4, fine-tuned for 2 epoch on training data with early stopping).

Table 5: Model Scores on test dataset: Highest scoring approaches are in bold.

Model	Accuracy	Precision	Recall	f1 score
Gradient Boosting	0.654	**0.691**	0.543	0.608
SVM	**0.668**	0.691	0.595	0.639
XGBoost	0.648	0.669	0.572	0.617
Random Forest	0.644	0.668	0.559	0.609
Llama 3.1*	0.662	0.645	0.708	**0.675**
BERT	0.644	0.639	0.646	0.643
AdaBoost	0.618	0.638	0.529	0.578
Logistic Regression	0.633	0.636	0.605	0.620
MLP	0.620	0.594	**0.729**	0.655
DecisionTree	0.573	0.566	0.590	0.578
Gemma 2*	0.543	0.543	0.477	0.507

*Zero-shot prompting approach used.

Model performance metrics, including precision, recall, and the F1 score, are reported for all models in Table 5 and visualised in Figure 2. The Gradient Boosting model demonstrated the highest precision, with a reported value of 0.691.

3.3. Ablation Study Generation

To assess feature importance in our feature-based classification models, we conducted an ablation study on all input string features (e.g., Title, Date

Published, Abstract). We created data sets for each feature by permuting the data to exclude that feature. We then calculated the average model evaluation metrics (F1 score, precision, recall, accuracy) across all models for each ablation. Lower scoring metrics indicate more contribution to a classifier's performance.

Table 6: Average Ablation Model Scores: lowest scoring ablation are in bold.

	Accuracy	Precision	Recall	f1 Score
Ablation				
Abstract Inverted Index	**0.629**	0.648	**0.556**	**0.597**
Citated By Count	0.635	0.654	0.565	0.605
First Author	0.633	0.648	0.574	0.607
First Author Countries	0.632	**0.644**	0.590	0.613
Primary Topic	0.633	0.651	0.560	0.601
Publication Date	0.630	0.645	0.570	0.604
Title	0.638	0.652	0.582	0.613

Averaged ablation study results are reported in Table 6 and visualised in Figure 3. Interestingly, within the ablation studies, the averaged lowest-scoring precision ablation was achieved by ablating First Author Countries (Precision 0.644). For all other metrics, the lowest averaged scored metric was achieved through ablating the Abstract (Accuracy 0.629, Recall 0.556, F1 Score 0.597).

3.4. Coefficient Analysis

To assess the importance of individual words in our classification task, we analysed the coefficients of our trained logistic regression model. We loaded the previously trained logistic regression model and the count vectorizer used to preprocess the text data. The feature names (individual words) of the count vectorizer were then extracted. These correspond to the columns in the document-term matrix used to train the model. The coefficients of the logistic regression model were extracted. In logistic regression, these coefficients represent the log-odds impact of each word on the classification decision. A positive coefficient indicates that the presence of the word increases the likelihood of a positive classification, while a negative coefficient decreases

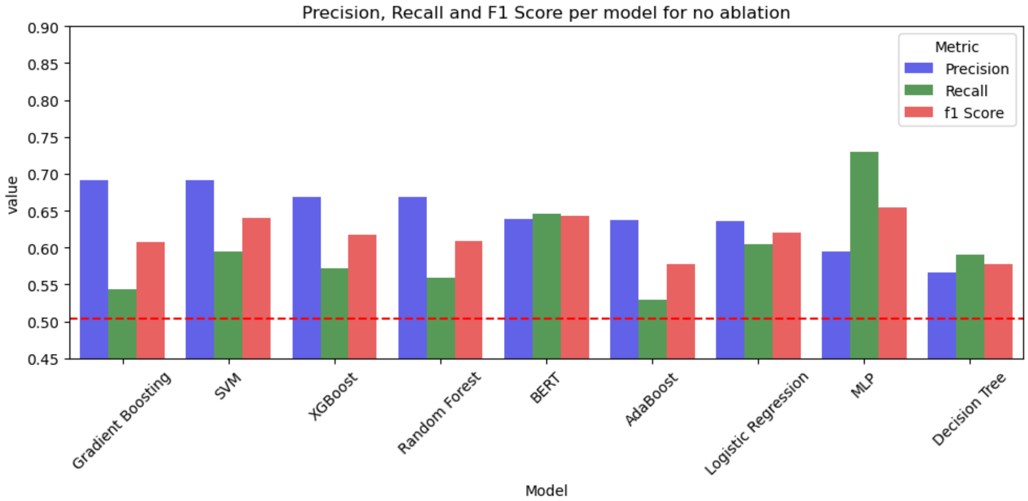

Figure 2: Different evaluation metrics on the test dataset.

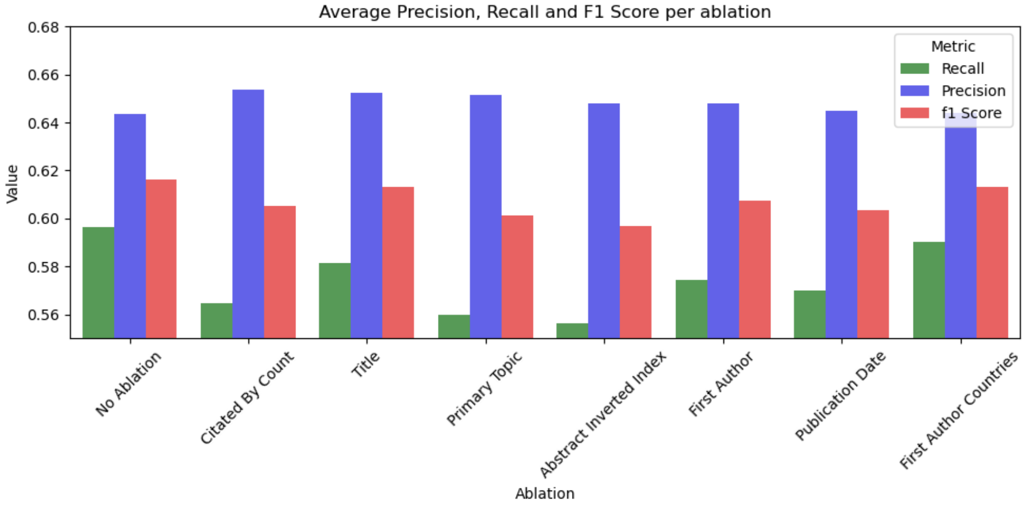

Figure 3: Average model scores per ablation study.

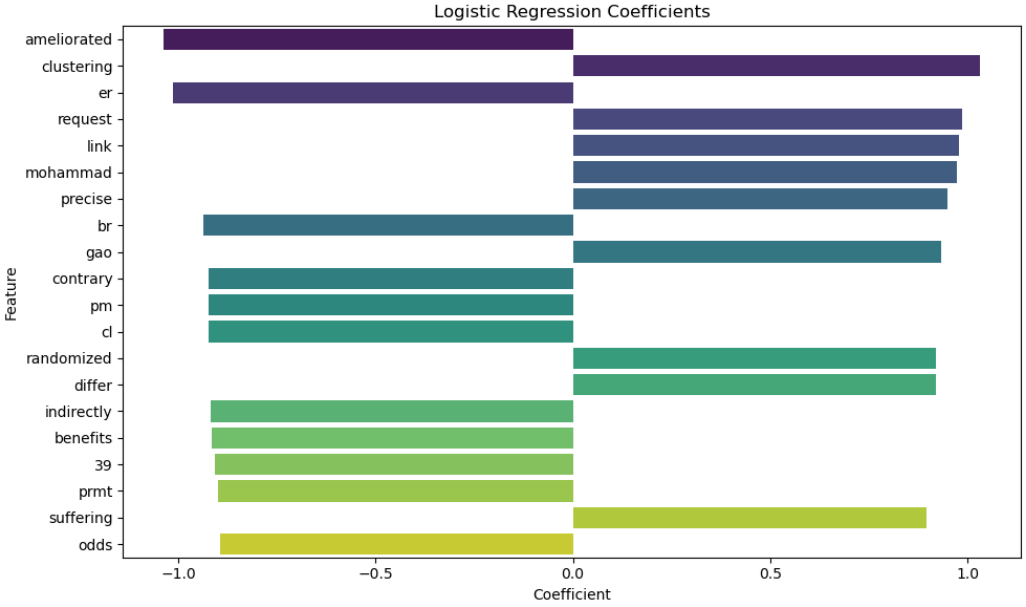

Figure 4: Coefficience visualisation of the linear regression model.

this likelihood. The absolute values of the coefficients were calculated to rank words according to their overall impact, regardless of direction (positive or negative). The features were then sorted in descending order of importance.

The coefficients are visualised in Figure 4.

4. Discussion

4.1. Journal Metadata analysis

Two notable findings were reported from the analysis of the journal metadata: a weak significant positive correlation between a journal's log of work found and the log of retraction count and a strong negative correlation between the journals' retraction count and log of a journal H index. This seems counterintuitive, as more retractions are likely to occur given more publications, and hence, a strong positive correlation would be present. This interesting finding could indicate that journals that publish fewer works are less proactive at detecting potential retractions or that publishing research that will be retracted is more complicated within journals with greater work output, presumably due to increased scrutiny of these works. The strong negative correlation between the retraction count and the log of the h-index indicates that as a journal's h-index increases, its retraction count tends to decrease significantly, which is expected.

The h-index, a widely used measure of a journal's productivity and impact, is based on its most cited papers. Typically when a work is considered for inclusion in a journal or conference, a peer reviewer is tasked with subjecting that research to the scrutiny of others who are experts in the same field [18]. This reviewer is sourced from academics who review for many, primarily, altruistic reasons, such as keeping up with the latest developments, building associations with journals, and demonstrating a commitment to the scientific field [19]. Importantly, not every researcher is a peer reviewer, which means that available review time is a finite resource [20]. It is likely that more reviwers are available for greater h-index journals. Publishing venues with higher h-index values potentially have a more rigorous peer review process, authors are more diligent when submitting to these journals, or higher-quality journals attract better-quality research.

4.2. Machine Learning For Predicting Retractions

This research demonstrates that machine learning techniques are appropriate for exploring and predicting article retractions. In particular, more traditional feature-based approaches such as gradient boost, SVM, XGBoost, and Random Forest achieved superior precision compared to the more modern contextually aware BERT model. However, this finding needs to be contextualised within the objectives of this investigation itself. The study aimed to establish baseline results for the generated data set rather than to optimise individual model performance. In particular, BERT, as a pre-trained model, typically requires extensive fine-tuning on large, domain-specific datasets to leverage its capabilities, which was not the case here. The limited fine-tuning described in the study (5 epochs on the training data) may have needed to have been sufficient to achieve superior precision performance in this domain. Furthermore, BERT was limited to 512 tokens, which could have truncated important information for this model, given the verbosity of the abstracts. The feature-based classifiers did not have this limitation. Furthermore, while BERT's precision was lower than traditional machine learning models, it demonstrated comparative performance in other metrics, such as recall and the F1 score. This finding could suggest that different model approaches could excel in different evaluation metrics.

Importantly, investigations into appropriate model selection lay the groundwork for future investigations. Optimising the performance of these approaches, particularly for specific metrics such as precision, remains an open challenge for this data set. Future work could explore more extensive fine-tuning of BERT and more elaborate feature engineering within feature-based classifier models.

4.3. Ablations

Several observations on the ablation of features can be made given the results reported in Table 6 and Figure 3. Unsurprisingly, given the amount of information contained within an abstract, it appears to be the most influential feature among all feature-based models when considering recall or F1 score, as when ablated, it resulted in the lowest average scores for accuracy (0.629), recall (0.556) and F1 score (0.597). This suggests that the abstract contains significant information to identify potentially retracted articles. Interestingly, ablating the ”First Author Countries” feature resulted in the lowest precision score (0.644). This indicates that the geographical origin of the first author provides valuable information for precise classification, which supports previous work outlined in the introduction.

4.4. Coefficient Analysis

Certain coefficients (words) were associated with the data set classes (retraction / non-retraction). Although analysis of this is speculative, the author suggests the reasons why certain coefficients were associated with classes as follows:

1. “Randomized” was strongly associated with a paper retraction; this could be due to the increased scrutiny that this type of research is subjected to (such as medical/health domains).
2. “Contrary”, which is likely to be seen in papers contradictory to established ideas, is less likely to have been retracted.
3. “Indirectly” and ”Benefits” have negative coefficients, which could suggest that more cautious or nuanced claims are less likely to be retracted.
4. The presence of particular names (“Mohammad” and “Gao”) could indicate some geographic or cultural factors in retractions, supporting previous research in this area [8].

4.5. Ethics of Automating Retraction Prediction

Using predictive models to identify potential retractions in the scientific literature raises several ethical concerns that warrant careful consideration.

Although these approaches offer promising tools for improving research integrity, they also present significant challenges that current methodologies have not adequately addressed. These concerns also partially explain why precision was the evaluation metric focused on when interpreting the results. A primary concern is that these models rely on correlations rather than causal relationships. This limitation may inadvertently perpetuate existing biases within the publication system. For instance, the coefficient analysis revealed an association between cautious language (e.g., "indirectly," "benefits") and retraction likelihood. However, such correlations do not necessarily imply causation and may lead to misinterpretation of results. Furthermore, the indiscriminate application of these models could potentially hinder the publication of genuinely innovative research that challenges established paradigms. Scientific progress often relies on work that contradicts prevailing doctrines, and we must be cautious not to impede such advancements. An examination of the model coefficients raises concerns about potential unintended consequences of using it's findings. Authors may be incentivised to engage in self-censorship or overly cautious reporting of results to avoid being flagged by these systems. Conversely, bad actors might exploit this knowledge to circumvent detection, potentially facilitating the dissemination of invalid results. Furthermore, implementing these models could introduce inductive bias into investigations, potentially leading to unforeseen consequences in the scientific publishing landscape. This bias can manifest itself in various ways, from shaping research questions to influencing methodological choices.

5. Conclusions

1. This research demonstrates that machine learning approaches can be used with some success to predict if an article is retracted.
2. Feature-based classifiers, such as gradient boosting machines and SVM, outperformed contextual approaches such as BERT.
3. The abstract of a work contains the most important feature for determining if a work is to be retracted (except for precision, where First Author's country is).

Appendix A. Appendix

Table A.7: Ablation Model Scores

Model	Ablation	Precision	Recall	f1 Score
Gradient Boosting	First Author Countries	0.691	0.543	0.608
SVM	First Author Countries	0.691	0.595	0.639
SVM	Citated By Count	0.691	0.595	0.639
SVM	Publication Date	0.689	0.580	0.630
Gradient Boosting	Title	0.688	0.532	0.600
SVM	First Author	0.685	0.591	0.635
Gradient Boosting	First Author	0.685	0.525	0.594
SVM	Primary Topic	0.684	0.597	0.637
Gradient Boosting	Publication Date	0.684	0.532	0.598
XGBoost	Title	0.683	0.570	0.621
Gradient Boosting	Citated By Count	0.682	0.531	0.597
Random Forest	Citated By Count	0.680	0.558	0.613
SVM	Abstract Inverted Index	0.680	0.571	0.621
Random Forest	Abstract Inverted Index	0.679	0.506	0.580
SVM	Title	0.678	0.593	0.632
MLP	Primary Topic	0.675	0.514	0.583
Gradient Boosting	Primary Topic	0.675	0.537	0.598
XGBoost	First Author	0.674	0.562	0.613
Random Forest	Title	0.673	0.567	0.616
Random Forest	Publication Date	0.672	0.559	0.610
XGBoost	Citated By Count	0.671	0.564	0.613
Random Forest	First Author	0.671	0.568	0.615
XGBoost	Primary Topic	0.669	0.562	0.611
Random Forest	Primary Topic	0.669	0.571	0.616
XGBoost	First Author Countries	0.669	0.572	0.617
Random Forest	First Author Countries	0.668	0.559	0.609
XGBoost	Publication Date	0.668	0.549	0.602
XGBoost	Abstract Inverted Index	0.666	0.555	0.605
MLP	Citated By Count	0.665	0.554	0.604
Gradient Boosting	Abstract Inverted Index	0.661	0.509	0.575
AdaBoost	Title	0.653	0.523	0.581
AdaBoost	Abstract Inverted Index	0.645	0.517	0.574
Logistic Regression	Citated By Count	0.639	0.606	0.622
AdaBoost	First Author Countries	0.638	0.529	0.578
Logistic Regression	Abstract Inverted Index	0.637	0.647	0.642
Logistic Regression	First Author Countries	0.636	0.605	0.620
AdaBoost	Citated By Count	0.635	0.528	0.576
AdaBoost	First Author	0.634	0.531	0.578
Logistic Regression	Primary Topic	0.632	0.606	0.619
AdaBoost	Publication Date	0.632	0.516	0.568
MLP	Abstract Inverted Index	0.631	0.597	0.613
Logistic Regression	First Author	0.629	0.606	0.618
AdaBoost	Primary Topic	0.629	0.514	0.566
MLP	Title	0.627	0.633	0.630
Logistic Regression	Title	0.626	0.616	0.621
Logistic Regression	Publication Date	0.625	0.603	0.614
MLP	First Author	0.623	0.625	0.624
MLP	Publication Date	0.618	0.644	0.631
MLP	First Author Countries	0.594	0.729	0.655
Decision Tree	Title	0.592	0.618	0.605
Decision Tree	Abstract Inverted Index	0.585	0.547	0.565
Decision Tree	First Author	0.581	0.587	0.584
Decision Tree	Primary Topic	0.579	0.578	0.578
Decision Tree	Publication Date	0.573	0.578	0.576
Decision Tree	Citated By Count	0.569	0.582	0.575
Decision Tree	First Author Countries	0.566	0.590	0.578

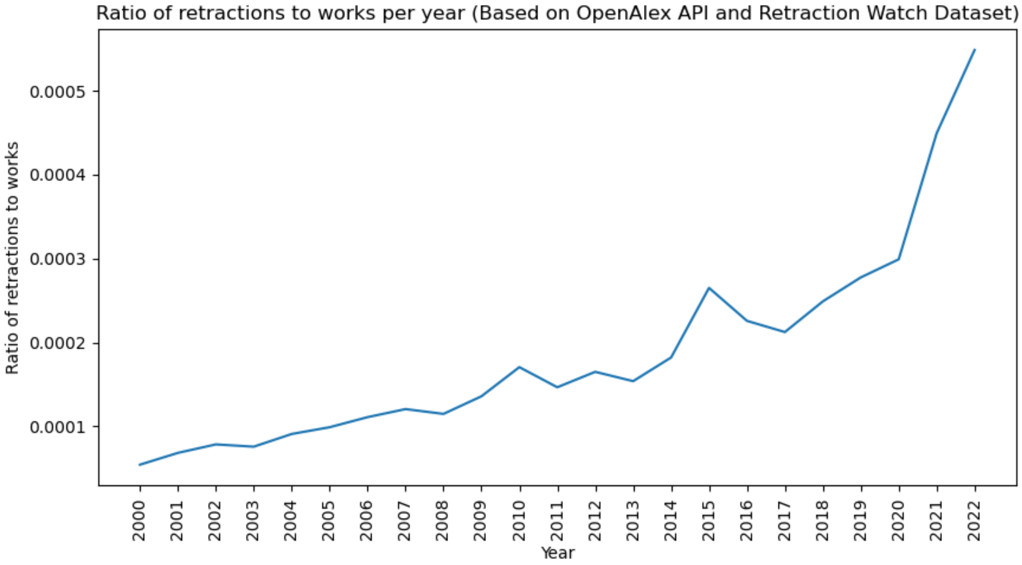

Figure A.5: Ratio of retractions to works: A positive increase is noted over the time period.

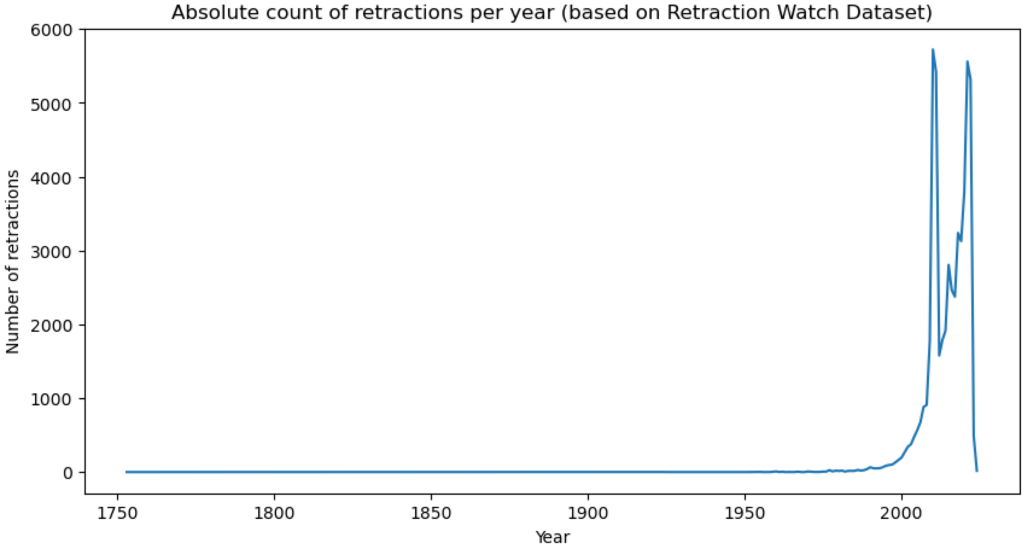

Figure A.6: Absolute count of retractions per year. Demonstrating that the absolute count of retractions has increased over the past 20 years.

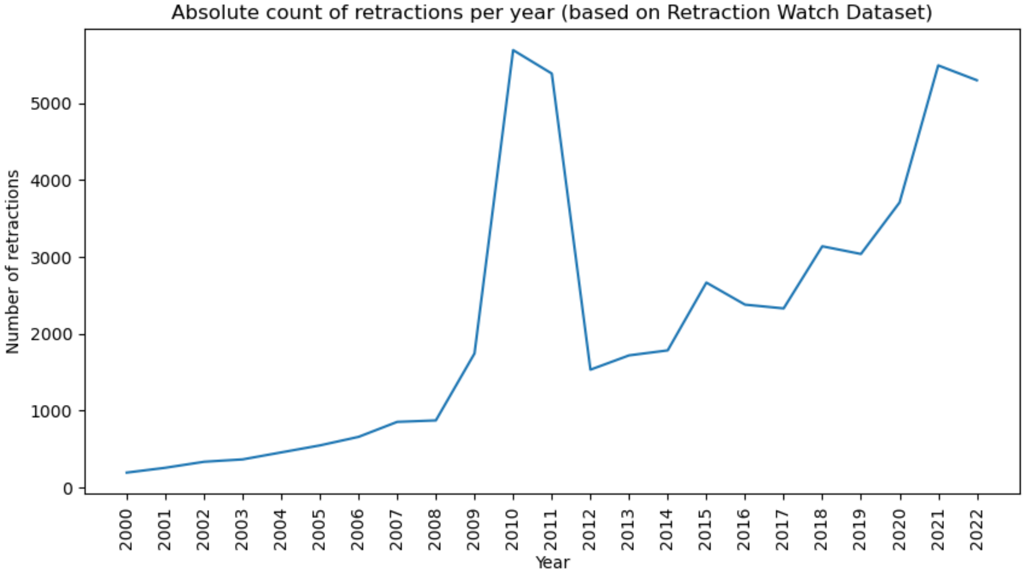

Figure A.7: Absolute count of retractions per year between 2000 and 2020. Note that peak of retractions around the 2010-2012 period, potentially due to the more than 8000 IEEE conference papers that were retracted in 2009-2011

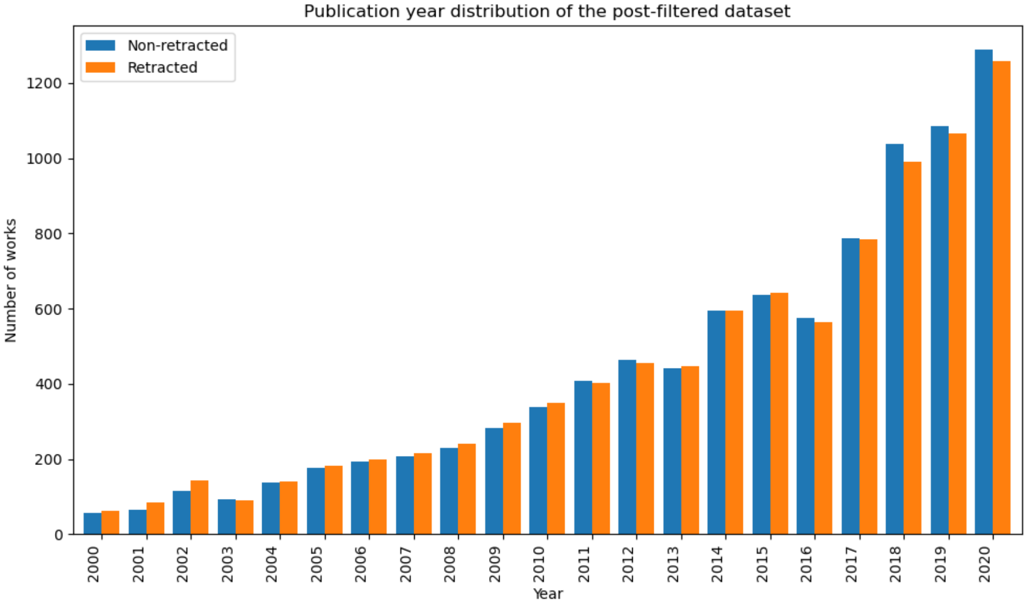

Figure A.8: Dataset publication parity: Publication year distribution for retracted and non-retracted works.

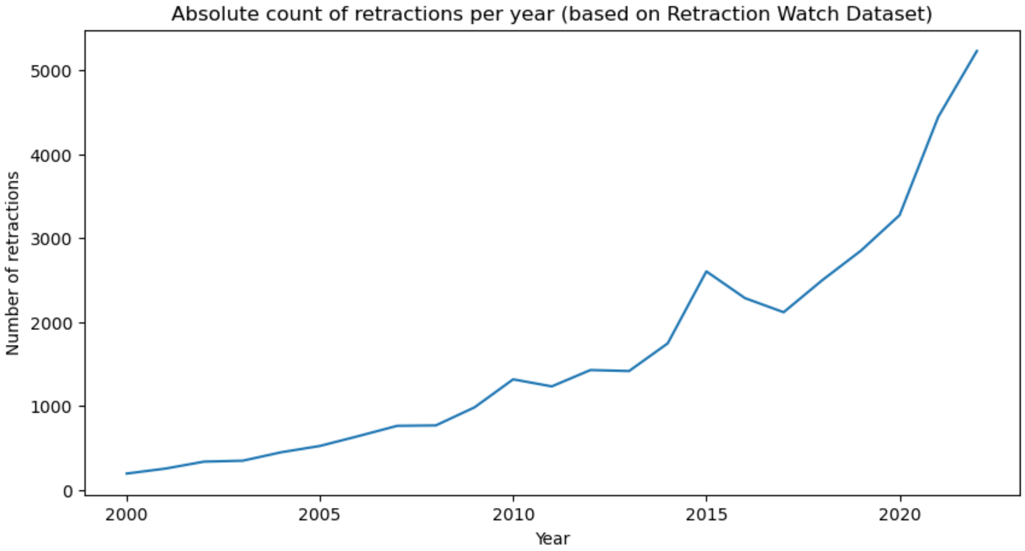

Figure A.9: Ratio of retracted works normalised by total works per year, after removing conference papers: An increasing trend for retractions.

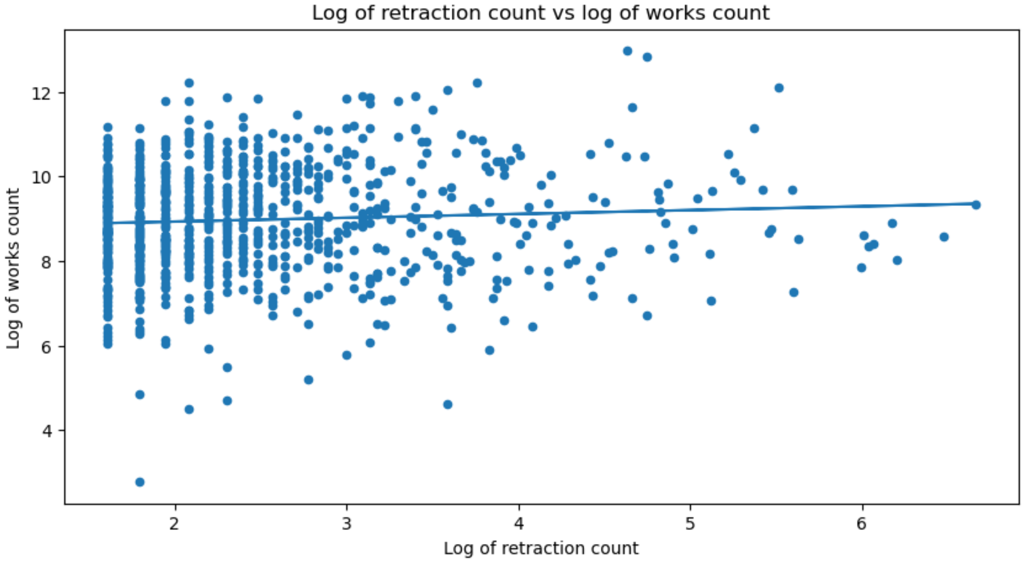

Figure A.10: A weak positive correlation between the log of works counts and log of retraction counts.

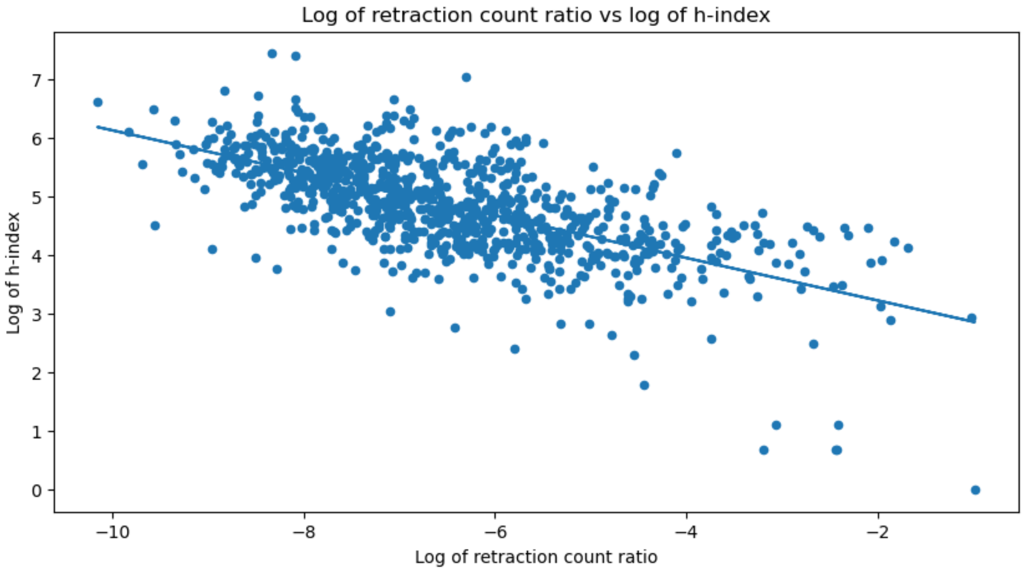

Figure A.11: A strong negative correlation between lof of h index and log of retraction counts.

References

[1] R. G. Steen, Retractions in the scientific literature: is the incidence of research fraud increasing?, Journal of Medical Ethics 37 (4) (2011) 249–253. `doi:10.1136/jme.2010.040923`.

[2] R. G. Steen, Retractions in the scientific literature: do authors deliberately commit research fraud?, Journal of Medical Ethics 37 (2) (2011) 113–117. `doi:10.1136/jme.2010.038125`.

[3] R. Perera, P. Nand, Recent Advances in Natural Language Generation: A Survey and Classification of the Empirical Literature, COMPUTING AND INFORMATICS 36 (1) (2017) 1–32, number: 1. `doi:https://doi.org/10.4149/cai_2017_1_1`.
URL `https://www.cai.sk/ojs/index.php/cai/article/view/2017_1_1`

[4] J. Dan, M. James H., Speech and Language Processing (3rd ed. draft), in: Speech and Language Processing (3rd ed. draft), Stanford University, 2024, pp. 1–23.

URL https://web.stanford.edu/~jurafsky/slp3/slides/4_NB_2024.pdf

[5] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, Deep Learning–based Text Classification: A Comprehensive Review, ACM Comput. Surv. 54 (3) (2021) 62:1–62:40. doi:10.1145/3439726.
URL https://doi.org/10.1145/3439726

[6] E. C. Moylan, M. K. Kowalczuk, Why articles are retracted: a retrospective cross-sectional study of retraction notices at BioMed Central, BMJ Open 6 (11) (2016) e012047, publisher: British Medical Journal Publishing Group Section: Ethics. doi:10.1136/bmjopen-2016-012047.
URL https://bmjopen.bmj.com/content/6/11/e012047

[7] E. R. Babbie, The Practice of Social Research, Cengage AU, 2020, google-Books-ID: KrGeygEACAAJ.

[8] S. Stretton, N. J. Bramich, J. R. Keys, J. A. Monk, J. A. Ely, C. Haley, M. J. Woolley, K. L. Woolley, Publication misconduct and plagiarism retractions: a systematic, retrospective study, Current Medical Research and Opinion 28 (10) (2012) 1575–1583. doi:10.1185/03007995.2012.728131.

[9] C. Candal-Pedreira, J. S. Ross, A. Ruano-Ravina, D. S. Egilman, E. Fernndez, M. Prez-Ros, Retracted papers originating from paper mills: cross sectional study, BMJ 379 (2022) e071517, publisher: British Medical Journal Publishing Group Section: Research. doi:10.1136/bmj-2022-071517.
URL https://www.bmj.com/content/379/bmj-2022-071517

[10] M. Gaudino, N. B. Robinson, K. Audisio, M. Rahouma, U. Benedetto, P. Kurlansky, S. E. Fremes, Trends and Characteristics of Retracted Articles in the Biomedical Literature, 1971 to 2020, JAMA Internal Medicine 181 (8) (2021) 1118–1121. doi:10.1001/jamainternmed.2021.1807.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8111562/

[11] R. Watch, Retraction Watch Database User Guide (Jun. 2024).
URL https://retractionwatch.com/wp-content/uploads/2023/12/Building-The-Database.pdf

[12] R. Watch, Retraction Watch (Jun. 2024).
URL https://retractionwatch.com/

[13] J. A. Teixeira da Silva, Silent or Stealth Retractions, the Dangerous Voices of the Unknown, Deleted Literature, Publishing Research Quarterly 32 (1) (2016) 44–53. doi:10.1007/s12109-015-9439-y.
URL https://doi.org/10.1007/s12109-015-9439-y

[14] J. Priem, H. Piwowar, R. Orr, OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, arXiv:2205.01833 [cs] (Jun. 2022). doi:10.48550/arXiv.2205.01833.
URL http://arxiv.org/abs/2205.01833

[15] R. Van Noorden, More than 10,000 research papers were retracted in 2023 a new record, Nature 624 (7992) (2023) 479–481, bandiera_abtest: a Cg_type: News Publisher: Nature Publishing Group Subject_term: Scientific community, Publishing. doi:10.1038/d41586-023-03974-8.
URL https://www.nature.com/articles/d41586-023-03974-8

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805 [cs] (May 2019). doi:10.48550/arXiv.1810.04805.
URL http://arxiv.org/abs/1810.04805

[17] F. Pedregosa, 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, . Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (85) (2011) 2825–2830.
URL http://jmlr.org/papers/v12/pedregosa11a.html

[18] D. Banks, Thoughts on Publishing the Research Article over the Centuries, Publications 6 (1) (2018) 10, number: 1 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/publications6010010.
URL https://www.mdpi.com/2304-6775/6/1/10

[19] P. J. Steer, S. Ernst, Peer review - Why, when and how, International Journal of Cardiology Congenital Heart Disease 2 (2021) 100083, publisher: Elsevier. doi:10.1016/j.ijcchd.2021.100083.

URL https://www.sciencedirect.com/science/article/pii/S2666668521000070

[20] V. Warne, Rewarding reviewers sense or sensibility? A Wiley study explained, Learned Publishing 29 (1) (2016) 41–50, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/leap.1002. doi:10.1002/leap.1002.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/leap.1002