Predicting retracted research

Aaron HA Fletcher[*] and Mark Stevenson[*]

[*]School of Computer Science, The University of Sheffield, Regent Court,
Sheffield, S1 4DP, United Kingdom.

Contributing authors: ahafletcher1@sheffield.ac.uk;
mark.stevenson@sheffield.ac.uk;

Abstract

Retracting published research is an important safeguard against the dissemination of flawed or fraudulent scientific information. This paper describes the creation of a novel dataset and machine learning models to predict which articles will be retracted. The dataset combines information from the Retraction Watch database and the OpenAlex API, including article metadata, abstracts, and citation metrics. A total of 9,028 articles (4,514 retracted and 4,514 non-retracted) published between 2000-20 were included. Several machine learning models were trained on this data, with a Support Vector Machine achieving the best precision for the recognition of retracted articles (0.690). The dataset and code are made publicly available to support future work in this area.

Keywords: Retraction prediction, Machine Learning, Scientific Publishing

1 Introduction

Retracting scientific articles is a crucial safeguard against disseminating inaccurate or unreliable information. Conversely, the number of journal retractions can act as a proxy for failures within the research dissemination 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 demonstrated increased retractions [1–4]. Obtaining definitive values for retraction numbers is difficult due to stealth retractions, where research is retracted without notification [5]. However, preventing the publication of inaccurate or unreliable information versus timely dissemination of research results remains a challenging balance [6].

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 for the authors [7] , the journal's reputation [8, 9] and the domain's integrity [10].

Many journals analyse submitted manuscripts using text similarity tools to identify plagiarised work that should not be re-published. However, direct plagiarism is one of many reasons a paper might subsequently be retracted. Automated methods to identify this work would benefit peer reviewers and journal editors in deciding which papers should be published. In addition, increasing amounts of research work are now distributed through non-peer-reviewed repositories such as arXiv [11], which may also benefit from automated analysis.

The decision over whether a piece of research should be retracted can be viewed as a binary classification task. Such tasks are common in natural language processing (NLP), where text is categorised into classes based on a particular aspect (such as sentiment, author or subject) [12, 13]. Recent advances in NLP, including word embeddings, deep neural networks, and transformer architectures, have demonstrated considerable success in text classification for a wide range of problems [14].

This paper explores the application of NLP methods to the problem of automatically identifying research that will subsequently be retracted. Its main contributions are to: report the development of a publicly available dataset of retracted and non-retracted research articles; apply this to train and evaluate a range of NLP-based classifiers designed to identify retracted articles.

2 Background

Failing to detect and address retracted research can significantly undermine the reliability of the scientific record. Avenell et al. demonstrated how just 12 retracted clinical trial reports—each tainted by misconduct—continued influencing scholarly discourse: they were cited by numerous publications, including systematic reviews and clinical guidelines [15]. Crucially, 13 of these reviews or guidelines would likely change their conclusions if the retracted reports were excluded, while eight others faced uncertain impacts. This underscores how even a handful of invalid findings can distort the broader evidence base.

Longitudinal studies have also highlighted the troubling persistence of citations to retracted articles. Schneider et al., for instance, examined the 11-year citation history of a 2008 retraction for falsified clinical trial data on omega-3 fatty acids [16]. They found that 96% of direct citations did not acknowledge the retraction, exacerbated by widespread inconsistencies in how retraction labels and metadata were handled across digital platforms. Similarly, Hsiao and Schneider analysed 7,813 retracted papers in PubMed and discovered that while citations gradually declined, retractions did not substantially change how these papers were cited; only 5.4% of citing contexts acknowledged the retraction [17]. These findings illustrate how inaccurate or misleading research can linger, especially when retraction notices are inconsistently applied.

High-profile cases remain particularly damaging, as demonstrated by Heibi and Peroni's analysis of the infamous paper linking the measles, mumps and rubella vaccine to autism [18]. Despite its retraction, citations to this paper continued to surge over two decades. Although some later discussions did acknowledge its retracted status, many did not delve into the specifics that invalidated its medical claims—highlighting the sustained influence of widely publicised but flawed research.

Interestingly, erroneous findings seem to not spread via indirect references. Van der Vet and Nijveen tracked a retracted, highly cited *Nature* paper and found no evidence that misinformation travelled through secondary citation chains [19]. Instead, flaws tended to appear only when subsequent authors cited the original paper directly. Nonetheless, across all these studies, a consistent theme emerges: once published, retracted research can continue shaping discourse unless it is clearly and repeatedly marked as invalid—underscoring the need for more robust and transparent retraction practices.

Determining the precise reasons for research retractions is a surprisingly complex and potentially unnecessary task for retraction classification. While retractions can generally be attributed to either honest errors or research misconduct (including data fabrication, plagiarism, or false authorship), evidence outlining the prevalence of each is inconsistent. For instance, Steen [2] found that 73.5% of PubMed retractions between 2000 and 2010 resulted from honest errors. In contrast, Moylan and Kowalczuk [20], in a study of BioMed Central journals, reported a significantly higher proportion (76%) of retractions due to misconduct. These discrepancies likely arise from the inherent difficulty in ascertaining researchers' true intentions, which leads to subjective and potentially unreliable classifications. Therefore, including potentially ambiguous retraction reasons as a feature for classification might hinder a model's ability to learn meaningful patterns within the data [21], suggesting that other, more objectively measurable features may be more effective. From a practical end-user standpoint, the 'why' behind a retraction may be of secondary importance compared to the 'what' — the crucial fact that the work is no longer considered valid. Therefore, the emphasis should be on accurate retraction identification, not on dissecting the often ambiguous and ultimately secondary issue of the reason. This underscores the critical need for robust methods to detect retracted research, regardless of the underlying cause.

Research into "paper mills", organisations that produce and sell fabricated manuscripts, offers another potential avenue for identifying indicators for research at high risk of retraction [22–24]. A 2022 study examining retracted medical articles using the Retraction Watch dataset, Web of Science, and journal citation reports highlighted a concerning increase in retractions linked to paper mills, with a notable concentration associated with China [25]. This study found the median time to retraction was two years, decreasing as the journal's impact factor increased. Further supporting the potential relevance of article type, 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 categorized as "meta" research [26]. Byrne [27] published practical guidelines for identifying paper mills based on the proposition that producing manuscripts at scale will likely result in shared textual features, organisational similarities, generic study hypotheses and experimental approaches, and manipulated images. These findings suggest that features such as an author's country of origin, institutional affiliation, and the type of article may contain valuable information for modelling retraction prediction. Starkly, research showing the organised and systematic nature of paper mill fraud reminds us of the potential significance and necessity of research into automated identification of retracted works.

The demographic patterns associated with paper mills, such as the overrepresentation of authors from specific countries, suggest that broader demographic factors potentially offer a signal for predicting retracted research. While research on the demographics of authors producing retracted articles remains limited, a significant association between first authors from lower-income countries and retractions due to plagiarism has been established, suggesting global variations in retraction reasons and the potential influence of socioeconomic factors [28]. Conversely, fraudulent research has been reported to be more prevalent in the United States, with over half (53%) of fraudulent articles authored by "repeat offenders" [1]. These contrasting findings underscore the issue's complexity and highlight the need for nuanced analysis that considers factors beyond basic demographic categories, such as institutional culture, research training, and publication pressures [24]. Nevertheless, they suggest that author-specific characteristics, such as an author's name (potentially indicative of origin), publication history, and institutional affiliation, could provide valuable information to assist in identifying research at higher risk of retraction.

We are unaware of any previous research investigating the use of machine learning to identify articles that will be retracted.

3 Methods/Experimental

The study aimed to develop and evaluate machine learning models, including feature-based classifiers and large language models (LLMs), for predicting the retraction status of academic articles. The research design used was a retrospective observational study with a case-control approach.

3.1 Dataset Construction

Publicly available data from two online databases were used to construct the dataset (Retraction Watch and OpenAlex)

Retraction Watch is a human-validated retraction dataset and is compiled from various sources, including journal databases, institutional reports, social media, and direct tips [29, 30]. Although not exhaustive due to unannounced or "stealth" retractions [5], it provides partial metadata for some retracted articles, such as title, journal, publisher, and author.

OpenAlex is open online catalogue of academic publications, similar to Scopus and Web of Science, which aggregates data from multiple sources and releases monthly updates. The combination of these resources was used to create a single dataset suitable for predicting article retractions [31].

A set of retracted articles were identified using the Retraction Watch dataset. Only articles and review works were considered. Conference papers were excluded due to a mass retraction of conference papers undertaken by the Institute of Electrical and

Table 1 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 2 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
Unretracted works's whose title contained	*"retraction", "retraction:", "withdrawn", "correction", "erratum", "retracted", "withdrawal", "conclusion", "editorial", "contributions", "commentary", "contributors".*
Retracted and unretracted works if abstract or title contained the words	*"elsevier", "notice", "editor", "editors", "publisher".*

Electronic Engineers between 2009 and 2011 (having retracted over 10,000 such papers in the past two decades) [32] and because there is no process to retract papers from many conference venues. Retractions were limited to a 20 year period from 2000 to 2020 due to the lack of retracted works before this date, the median post-publication time to retraction being 1.8 years [26] and the increased use of natural language technologies subsequent to this period. Information from OpenAlex (API queried on 24/07/2024) was also used to filter out some works. A full list of exclusion criteria for journals and articles are shown in Tables 1 and 2.

For each retracted article, another article was randomly sampled that was published in the same year as the retracted article, did not meet the works exclusion

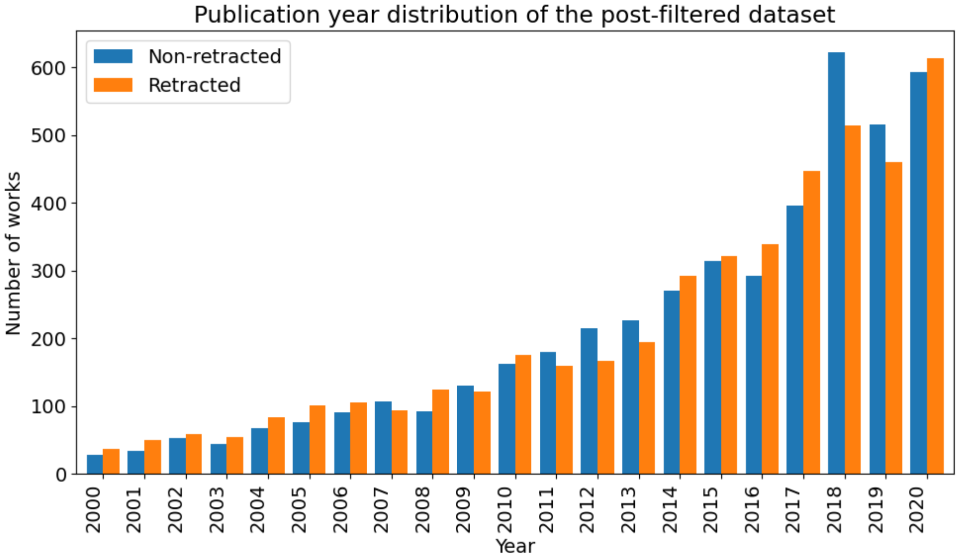

Fig. 1 Publication year distribution for retracted and non-retracted works.

criteria outlined in Table 2, was not included in the retraction watch dataset and whose OpenAlex API flag of *'is_retracted'* was False.

Articles containing keywords strongly indicating that it has been retracted were also excluded, using the following list of keywords: "retraction", "retracted", "retract", "retractionwatch", "retraction watch", "removed", "withdrawn", "withdrawal", "withdraw", "retracted article" and "article".

The following features were extracted for each work (retracted and non-retracted): Abstract Inverted Index, Publication Date, Primary Topic, First Author, Institution, Citation Count, First Author Countries, Is Retracted Flag and Article Type.

3.2 Dataset Characteristics

The dataset was balanced through undersampling, resulting in a total of 9,028 pieces of research, with equal numbers of retracted and non-retracted. It was divided into training (64%), validation (16%) and test sets (20%).

The distribution of works in each year is shown in Figure 1. It can be seen that the number of included works associated with each year increases over time, reflecting the trend of retractions increasing over time in the original Retraction Watch dataset.

In the generated dataset, 7.54% of the articles reported as retracted in Retraction Watch were not marked as retracted by OpenAlex, possibly because OpenAlex's metadata is derived from multiple input sources. This discrepancy further illustrates the difficulty of identifying retracted research since it may not be labelled as such.

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). This seems counterintuitive, as more retractions are likely to occur given more publications, and hence, a strong positive correlation would be present. This 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.
2. A strong negative correlation between the retraction count and log of h-index (Pearson correlation coefficient -0.656, p-value < 0.05). This relationship 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 [33]. This reviewer is sourced from academics who review for many reasons (primarily altruistic), such as keeping up with the latest developments, building associations with journals, and demonstrating a commitment to the scientific field [34]. Importantly, time available to review is a finite resource [35]. It is likely that more reviewers 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.

3.3 Retraction Classifier

A range of approaches to predict retractions were explored using feature-based classifiers and LLMs. A range of feature-based classifiers that have been demonstrated to be effective for text-classification tasks were used: Gradient Boosting, Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost), Random Forest, Mutli-layer Perceptron (MLP) and Decision Trees [36, 37]. A Super Learner model, an ensemble approach with multiple machine learning models, was also utilised [32].

A common set of features was shared by these classifiers: Title + Abstract, Primary Topic, First Author, Publication Year, and Citation Count. All textual fields (Title, Abstract, Primary Topic and First Author) were converted to lowercase with non-ASCII characters, special punctuation, and numbers removed. The title and abstract were then combined. The Publication Date feature was converted solely to its year in YYYY format. Numerical features (Publication Year and Citation Count) were min-max normalised to scale between 0 and 1. Categorical features (First Author's country) were one-hot encoded based on the training data subset. All other text features were represented using term weights produced by Best Matching 25 [38] with $k = 2$, $b = 0.3$ and no maximum vocabulary length specified.

LLMs have recently been demonstrated to be capable of high performance in a wide range of NLP tasks. A range of LLMs models were used, including models pre-trained on generalised data sources, i.e. Large Language Model Meta AI (Llama) 3.2 ("unsloth/Llama-3.2-3B-bnb-4bit"), its instruct variant ("unsloth/Llama-3.2-3B-Instruct-bnb-4bit") and Bidirectional Encoder Representations from Transformers

(BERT) ("bert-base-uncased"), pre-trained on a target domain, i.e. BioBERT ("dmis-lab/biobert-base-cased-v1.2") and recent high performing models, i.e. Gemma 2 base ("unsloth/gemma-2-9b") and it's instruct variant ("unsloth/gemma-2-9b-it-bnb-4bit"). LLM fine-tuning was conducted using supervised fine-tuning for a maximum of 10 epochs, with early stopping implemented to prevent overfitting.

Inputs for the encoder-based LLM models (BERT and BioBERT) consisted of the features (Text, Primary Topic, First Author, First Author Country, Citated By Count and Publication Year) concatenated with [SEP] tokens separating each feature - as shown in Figure 2. Only the pooling layer and the classification head could be trained for these encoder-based models. Decoder-based LLMs, that is, Llama and Gemma models, were fine-tuned using the input format and prompt template illustrated in Figures 3 and 4. A low-order rank adaptation approach was used during fine-tuning, with output vocabulary restricted to yes and no tokens. During testing, these models were evaluated by providing the input text and the question without the label to assess their ability to predict the retraction status independently, with a softmax of the logits for the yes and no output tokens forming the model's prediction. Commercially available LLMs were evaluated (GPT-4o mini [39] and Claude 3.5 sonnet [40]) using a zero-shot approach. These models were queried using the same prompting template as the fine-tuned models, but without any task-specific fine-tuning. All commercial models responded that no research was retracted within the testing dataset. This was thought to be due to the safety restrictions implemented within these models, which prevented responses that could be considered problematic [41].

INPUT_TEXT Formation

```
{title + abstract} [SEP] {primary topic}
[SEP] {first author}  [SEP] {first author country}
[SEP] {citated by count} [SEP] {publication year}
```

Fig. 2 Input format for encoder-based models.

INPUT_TEXT Formation

```
Text: {title + abstract} Primary Topic: {primary topic}
First Author: {first author}
First author country: {first author country}
Citated by count: {citated by count}
Publication Year: {publication year}
```

Fig. 3 Input format for decoder-based models.

Prompt Template

```
Here is a research article:
{INPUT_TEXT}

Is this paper retracted?
The correct answer is: {LABEL}
```

Fig. 4 Prompt template used for structuring input data during model fine-tuning.

All models were trained on the training dataset and evaluated on the test dataset, with the LLMs early stopping being determined using the validation dataset. LLMs were fine-tuned using pre-trained weights.

Model performance is measured using standard metrics for classification problems. Accuracy is the proportion of instances correctly classified as either retracted or non-retracted. Precision, recall and F1 scores are computed individually for the retracted and non-retracted classes and then averaged.

4 Results

4.1 Classifier Performance

Results for all classifiers are presented in Table 3 which shows performance for both the retracted and not retracted classes. (Overall classification performance scores can be obtained by averaging figures across classes.) The highest-scoring approaches for each metric are highlighted in bold.

All models outperformed random guessing (0.5 for all metrics), although the improvement varies considerably between models. The highest accuracy (0.682) is achieved by Llama 3.2-base, although accuracy scores overall are generally higher for more traditional feature-based approaches such as gradient boost, SVM, XGBoost, and Random Forest achieved superior precision compared to the more modern contextually aware LLMs.

Regarding the retracted class, SVM achieved the highest precision (0.690) and Llama 3.2-base the highest recall (0.683). Interestingly, both instruction-tuned decoder-based LLMs (Gemma 2-instruct and Llama 3.2-instruct) also achieve high recall for the retracted class but this is achieved by predicting retracted for the majority of instances, as demonstrated by the very low recall for the non-retracted class. This could be due to an effect of instruction tuning, as they are trained to be more cautious and risk-averse, indicating that instruction-tuned models might not be suitable for this type of classification task.

These findings establish baseline results using the dataset.

4.2 Ablation Study

The importance of individual features to the feature-based classification models was explored by conducting an ablation study on all input features. Datasets were created

Table 3 Retraction classifier performance results.

Model	Acc.	Non-Retracted			Retracted		
		P	R	F1	P	R	F1
Logistic Regression	0.638	0.638	0.647	0.642	0.639	0.630	0.635
Decision Tree	0.568	0.570	0.574	0.572	0.568	0.564	0.566
Random Forest	0.666	0.648	**0.731**	**0.687**	0.689	0.601	0.642
SVM	0.671	0.655	0.725	0.688	**0.690**	0.616	0.651
XGBoost	0.665	0.654	0.705	0.679	0.678	0.624	0.650
AdaBoost	0.631	0.619	0.684	0.650	0.645	0.577	0.609
Super Learner	0.669	0.661	0.699	0.680	0.678	0.640	0.659
MLP	0.655	0.650	0.675	0.663	0.660	0.634	0.647
Gemma 2-base	0.553	0.615	0.292	0.396	0.534	0.816	0.645
Gemma 2-instruct	0.529	**0.730**	0.098	0.173	0.515	**0.963**	0.671
BERT	0.609	0.612	0.602	0.607	0.606	0.616	0.611
BioBERT	0.608	0.598	0.668	0.631	0.621	0.548	0.582
Llama 3.2-base	**0.682**	0.686	0.674	0.680	0.678	0.689	**0.683**
Llama 3.2-instruct	0.535	0.714	0.121	0.208	0.518	0.951	0.671

for each feature by permuting the data to exclude that feature and then averaging the evaluation metrics (F1 score, precision, recall, accuracy) across all models for each ablation. Lower scoring metrics indicate a greater contribution to the performance of a classifier.

Table 4 Ablation performance metrics: lowest scoring ablations are in bold.

Model	Acc.	Non-Retracted			Retracted		
		P	R	F1	P	R	F1
Abstract	0.655	0.670	0.612	0.638	0.645	0.678	0.669
Citation Count	0.648	0.659	0.611	0.634	0.638	0.684	0.660
First Author	0.649	0.662	0.609	0.634	0.639	0.689	0.663
First Author Countries	0.649	0.663	0.604	0.632	0.638	0.693	0.664
Primary Topic	0.644	0.658	0.602	0.628	0.634	0.687	0.659
Publication Year	**0.641**	**0.656**	**0.594**	**0.622**	**0.630**	**0.688**	**0.657**
Title	0.648	0.657	0.618	0.636	0.641	0.678	0.658

Several observations on the ablation of features can be made given the results reported in Table 4. The publication year proved to be the most crucial feature, with its ablation resulting in the lowest scores across all metrics. This suggests that temporal information plays a more significant role in classification than detailed textual content, which is logical given the increase in publications and the corresponding increase in retractions. This temporal information component also highlights the challenge that traditional classification approaches potentially face in this area; given that publication year is such a strong signal, its presence might eclipse other valuable contributions from other features. The Primary Topic feature also demonstrated substantial importance, producing the second-lowest scores when ablated. Reduction in performance when First Author Countries are ablated provides some indication of the likelihood that a work will be retracted, supporting previous findings [28].

Contrary to what might be intuitively expected, the abstract, despite being the longest and most detailed textual component, emerged as the least influential feature across all evaluation metrics. When ablated, it yielded the highest average scores for accuracy (0.655), precision (0.657), recall (0.655), and F1 score (0.654), indicating its removal had the least negative impact on model performance. This counterintuitive finding regarding the abstract's limited influence could be attributed to several factors. First, structured metadata features (like publication date and primary topic) may provide more consistent and unambiguous signals for classification compared to the potentially noisy and variable nature of abstract text. Second, there might be considerable information redundancy between the abstract and other textual features like the title, making its individual contribution less distinctive.

5 Discussion

One of the potential applications of the classifier described above is as a tool during the peer review process, in much the same way that text similarity tools are often used to identify potential plagiarism. The required level of precision or recall would depend on how the tools would be used. If used as a screening tool to flag potentially problematic papers for additional review, a high recall would be preferable to avoid missing articles that are subsequently retracted. However, if used as a check which a submitted article must pass then high precision would be necessary to avoid the suppression of valid research. The performance of the models reported above, while promising, indicates that identification of retracted articles is not a trivial prediction task and may not be sufficient for some purposes. The decision regarding the involvement of systems to detect potential retractions within the peer review process is ultimately the choice of publishers.

Automatic prediction of potential retractions also raises ethical concern. Predictive models, such as the ones described here, have the potential to introduce bias thereby raising potential fairness issues [42, 43]. Such biases have the potential to unfairly penalise the groups who are more likely to be identified as producing research that will be retracted (e.g. first authors from particular locations) while benefiting those it is less likely to identify. This could introduce inductive bias into investigations, potentially leading to unforeseen consequences in the scientific publishing landscape such as influencing which research questions are investigated and which methodologies applied. In addition, authors may attempt to report results in ways that avoid detection by these models, potentially leading to self-censorship or overly cautious reporting of results. Conversely, bad actors with knowledge of these models may exploit that information to avoid detection, potentially facilitating the dissemination of invalid results.

6 Conclusions

This research demonstrates the potential of machine learning approaches in predicting retracted articles, contributing to efforts aimed at enhancing the integrity of scientific publication. By creating a novel open source dataset that combines information from the Retraction Watch database and the OpenAlex API, a resource for future investigations in this area has been contributed. Our dataset encompasses 16,224 articles published between 2000 and 2020, evenly divided between retracted and non-retracted works, and includes a variety of features such as abstracts, citation metrics, and author information.

Experiments showed that, with the exception of the recently released Llama 3.2 base model, traditional feature-based classifiers, such as gradient boosting machines and SVMs, outperformed contextual language models like BERT, BioBERT, and Gemma in terms of precision. The best-performing model achieved a precision of 0.690, indicating that while machine learning techniques hold promise, there remains a need for significant improvement before they can be effectively integrated into the peer review process. The ablation study highlighted the importance of the publication year, primary topic and the first author's country in predicting retractions, aligning with previous findings that suggest certain demographics may be more prone to retractions due to various factors.

6.1 Future work

There is potential for the approaches described here to be extended by making use of additional information with the potential to assist in the identification of retracted research. For example, the citation network of references to a paper and the references within the paper itself may provide useful information. In addition, the models described here analysed abstracts, but analysis of the full text itself could potentially allow models to evaluate flaws in methodology, result synthesis or false conclusions. Finally, analysis of the full author list of an article could reveal patterns of collaboration or even help to identify potential paper mills.

7 List of Abbreviations

- NLP: Natural Language Processing
- SVM: Support Vector Machine
- MLP: Multilayer Perceptron
- LLM: Large Language Machine
- XGBoost: Extreme Gradiant Boosting
- ASCII: American standard code for information exchange
- BERT: Bidirectional Encoder Representations from Transformers
- LLAMA: Large Language Model Meta AI

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The dataset supporting the conclusions of this article is available in the Predicting Article Retractions repository [44].

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

Authors' contributions

AF contributed to this research's conception, design analysis, data interpretation, and submission drafting. MS contributed to this research's conception, design analysis, data interpretation, and submission drafting. All authors read and approved the final manuscript.

Open Access

References

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

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

[3] Steen, R.G., Casadevall, A., Fang, F.C.: Why has the number of scientific retractions increased? PLoS ONE **8**(7), 68397 https://doi.org/10.1371/journal.pone.0068397

[4] Kühberger, A., Streit, D., Scherndl, T.: Self-correction in science: The effect of retraction on the frequency of citations. PloS One **17**(12), 0277814 https://doi.org/10.1371/journal.pone.0277814

[5] Teixeira Da Silva, J.A.: Silent or stealth retractions, the dangerous voices of the unknown, deleted literature. Publishing Research Quarterly **32**(1), 44–53 https://doi.org/10.1007/s12109-015-9439-y

[6] Perera, R., Nand, P.: Recent Advances in Natural Language Generation: A Survey and Classification of the Empirical Literature. Computing and Informatics **36**(1), 1–32 (2017) https://doi.org/10.4149/cai_2017_1_1

[7] Kearney, M., Downing, M., Gignac, E.A.: Research integrity and academic medicine: the pressure to publish and research misconduct. Journal of Osteopathic Medicine **124**(5), 187–194 (2024) https://doi.org/10.1515/jom-2023-0211

[8] Caporale, C., Zagarella, R.M.: Ethics and Integrity in Academic Publishing. In: Congiunti, L., Lo Piccolo, F., Russo, A., Serio, M. (eds.) Ethics in Research, pp. 53–69. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-24060-7_5 . Series Title: UNIPA Springer Series. https://link.springer.com/10.1007/978-3-031-24060-7_5 Accessed 2025-01-24

[9] Collaborative Working Group from the conference "Keeping the Pool Clean: Prevention and Management of Misconduct Related Retractions": RePAIR consensus guidelines: Responsibilities of Publishers, Agencies, Institutions, and Researchers in protecting the integrity of the research record. Research Integrity and Peer Review **3**(1), 15 (2018) https://doi.org/10.1186/s41073-018-0055-1 . Accessed 2025-01-24

[10] Grey, A., Avenell, A., Klein, A.A., Byrne, J.A., Wilmshurst, P., Bolland, M.J.: How to improve assessments of publication integrity. Nature **632**(8023), 26–28 (2024) https://doi.org/10.1038/d41586-024-02449-8

[11] arXiv.org e-Print Archive. https://arxiv.org/ Accessed 2025-01-15

[12] Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edn., pp. 1–23 (2025). Online manuscript released January 12, 2025. https://web.stanford.edu/~jurafsky/slp3/ Accessed 2025-1-12

[13] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep Learning–based Text Classification: A Comprehensive Review. ACM Comput. Surv. **54**(3), 62–16240 (2021) https://doi.org/10.1145/3439726

[14] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv (2019). https://doi.org/10.48550/arXiv.1810.04805

[15] Avenell, A., Stewart, F., Grey, A., Gamble, G., Bolland, M.: An investigation into the impact and implications of published papers from retracted research: systematic search of affected literature. BMJ Open **9**(10) (2019) https://doi.org/

10.1136/bmjopen-2019-031909

[16] Schneider, J., Ye, D., Hill, A.M., Whitehorn, A.S.: Continued post-retraction citation of a fraudulent clinical trial report, 11 years after it was retracted for falsifying data. Scientometrics **125**(3), 2877–2913 (2020) https://doi.org/10.1007/s11192-020-03631-1

[17] Hsiao, T.-K., Schneider, J.: Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions shown in citation contexts in biomedicine. Quantitative Science Studies **2**(4), 1144–1169 (2021) https://doi.org/10.1162/qss_a_00155

[18] Heibi, I., Peroni, S.: A qualitative and quantitative analysis of open citations to retracted articles: the wakefield 1998 et al.'s case. Scientometrics **126**(10), 8433–8470 (2021) https://doi.org/10.1007/s11192-021-04097-5

[19] Vet, P.E., Nijveen, H.: Propagation of errors in citation networks: a study involving the entire citation network of a widely cited paper published in, and later retracted from, the journal nature. Research Integrity and Peer Review **1**(1), 3 (2016) https://doi.org/10.1186/s41073-016-0008-5

[20] Moylan, E.C., Kowalczuk, M.K.: Why articles are retracted: a retrospective cross-sectional study of retraction notices at biomed central. BMJ Open **6**(11) (2016) https://doi.org/10.1136/bmjopen-2016-012047

[21] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2nd edn. Springer Series in Statistics. Springer. https://doi.org/10.1007/978-0-387-84858-7

[22] Hvistendahl, M.: China's publication bazaar. Science **342**(6162), 1035–1039 (2013) https://doi.org/10.1126/science.342.6162.1035

[23] Liu, X., Chen, X.: Journal retractions: Some unique features of research misconduct in china. Journal of Scholarly Publishing **49**(3), 305–319 (2018) https://doi.org/10.3138/jsp.49.3.02

[24] Tian, M., Su, Y., Ru, X.: Perish or publish in china: Pressures on young chinese scholars to publish in internationally indexed journals. Publications **4**(2) (2016) https://doi.org/10.3390/publications4020009

[25] Candal-Pedreira, C., Ross, J.S., Ruano-Ravina, A., Egilman, D.S., Fernández, E., Pérez-Ríos, M.: Retracted papers originating from paper mills: cross sectional study. BMJ **379**, 071517 (2022) https://doi.org/10.1136/bmj-2022-071517

[26] Gaudino, M., Robinson, N.B., Audisio, K., Rahouma, M., Benedetto, U., Kurlansky, P., Fremes, S.E.: Trends and Characteristics of Retracted Articles in the Biomedical Literature, 1971 to 2020. JAMA Internal Medicine **181**(8), 1118–1121

(2021) https://doi.org/10.1001/jamainternmed.2021.1807

[27] Byrne, J.A., Christopher, J.: Digital magic, or the dark arts of the 21st century—how can journals and peer reviewers detect manuscripts and publications from paper mills? FEBS Letters **594**(4), 583–589 (2020) https://doi.org/10.1002/1873-3468.13747

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

[29] Retraction Watch. The Center for Scientific Integrity, New York (2018). https://retractionwatch.com/ Accessed 2024-06-14

[30] Retraction Watch Database User Guide (2024). https://retractionwatch.com/wp-content/uploads/2023/12/Building-The-Database.pdf Accessed 2024-06-14

[31] Priem, J., Piwowar, H., Orr, R.: OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv (2022). https://doi.org/10.48550/arXiv.2205.01833

[32] Van Noorden, R.: More than 10,000 research papers were retracted in 2023 — a new record. Nature **624**(7992), 479–481 (2023) https://doi.org/10.1038/d41586-023-03974-8

[33] Banks, D.: Thoughts on Publishing the Research Article over the Centuries. Publications **6**(1), 10 (2018) https://doi.org/10.3390/publications6010010

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

[35] Warne, V.: Rewarding reviewers – sense or sensibility? A Wiley study explained. Learned Publishing **29**(1), 41–50 (2016) https://doi.org/10.1002/leap.1002

[36] Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**(1), 1–47 (2002) https://doi.org/10.1145/505282.505283

[37] Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P.S., He, L.: A survey on text classification: From traditional to deep learning. ACM Trans. Intell. Syst. Technol. **13**(2) (2022) https://doi.org/10.1145/3495162

[38] Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr. **3**(4), 333–389 (2009) https://doi.org/10.1561/1500000019

[39] OpenAI Platform. https://platform.openai.com Accessed 1/12/2024

[40] Introducing Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet Accessed 11/12/2024

[41] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S.E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S.R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., Kaplan, J.: Constitutional AI: Harmlessness from AI Feedback (2022). https://arxiv.org/abs/2212.08073

[42] Caton, S., Haas, C.: Fairness in machine learning: A survey. ACM Comput. Surv. **56**(7) (2024) https://doi.org/10.1145/3616865

[43] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6) (2021) https://doi.org/10.1145/3457607

[44] Anonymised Repository - Anonymous GitHub. https://anonymous.4open.science/r/RetractionWatch/README.md Accessed 2025-01-15