Analyzing the Evolution of Scientific Misconduct Based on the Language Of Retracted Papers

Christof Bless[1,2 *], **Andreas Waldis**[1,4], **Angelina Parfenova**[1,3],
Maria Andueza Rodriguez[1,5], **Andreas Marfurt**[1]
[1]Lucerne University of Applied Sciences and Arts,
[2]Leibniz University Hannover,
[3]Technical University of Munich,
[4]Technical University of Darmstadt,
[5]University of Fribourg

Abstract

Amid rising numbers of organizations producing counterfeit scholarly articles, it is important to quantify the prevalence of scientific misconduct. We assess the feasibility of automated text-based methods to determine the rate of scientific misconduct by analyzing linguistic differences between retracted and non-retracted papers. We find that retracted works show distinct phrase patterns and higher word repetition. Motivated by this, we evaluate two misconduct detection methods, a mixture distribution approach and a Transformer-based one. The best models achieve high accuracy (>0.9 F1) on detection of paper mill articles and automatically generated content, making them viable tools for flagging papers for closer review. We apply the classifiers to more than 300,000 paper abstracts, to quantify misconduct over time and find that our estimation methods accurately reproduce trends observed in the real data.

1 Introduction

The integrity of scientific research is increasingly threatened by the rise of so-called *paper mills*, for-profit organizations that produce and sell fraudulent academic manuscripts to researchers, academics, or students who are under pressure to publish in peer-reviewed journals (Candal-Pedreira et al., 2022; Abalkina, 2023). Often disguised as editing or translation services, paper mills sell manuscripts, author slots on peer-reviewed papers, and citations for existing papers (COPE, 2025; Christopher, 2021). Papers produced by paper mills can have negative consequences for society as they circulate false claims, erode trust in science, or lead to unjustified academic promotions (Byrne et al., 2022; Fanelli et al., 2021).

Since 2010, at least 5402 retracted papers have been connected to paper mills according to the Retraction Watch database (The Center for Scientific Integrity, 2018). This would be around 2 in every 10,000 papers indexed by Scopus in the same time frame. However, this number reflects only cases where the paper mill activity was uncovered. The real number of fabricated papers is likely considerably higher due to their convincing nature (Oransky et al., 2021; Brainard and You, 2018).

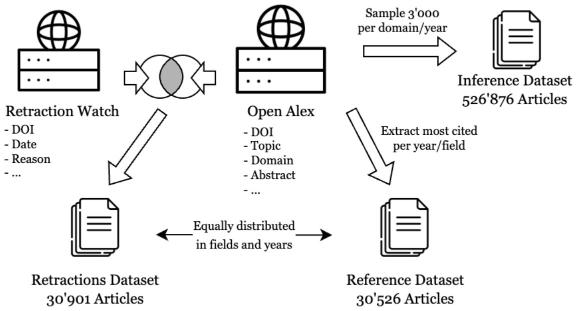

Retraction Watch
- DOI
- Date
- Reason
- ...
Open Alex
- DOI
- Topic
- Domain
- Abstract
- ...
Sample 3'000 per domain/year
Inference Dataset 526'876 Articles
Extract most cited per year/field
Equally distributed in fields and years
Retractions Dataset 30'901 Articles
Reference Dataset 30'526 Articles

Figure 1: We create a dataset of retracted papers by merging resources from `retractionwatch.org` and `openalex.org`. We add to that a reference dataset of non-retracted articles to train various classifiers for identifying papers with scientific misconduct. Finally, we estimate the development of misconduct over time by running a classifier on a large inference dataset of papers from diverse domains.

In this paper, we evaluate the feasibility of estimating the true rate of scientific misconduct by analyzing linguistic differences between retracted and non-retracted papers and specifically papers retracted for reasons of scientific misconduct. Based on preliminary data review, the central hypothesis is that paper mill articles share a distinctive writing style characterized by words and phrases stemming from the methods used to produce these articles. We conjecture that the automated tools and/or human ghostwriters share a common style or produce unusual expressions (Cabanac et al., 2021).

To investigate this hypothesis, we construct a text corpus of retracted and non-retracted articles (Figure 1 and Section 2) and assess linguistic characteristics of retracted papers (Section 3). Next,

[*] Corresponding author: christof.bless@hslu.ch

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 57–71
July 31, 2025 ©2025 Association for Computational Linguistics

we evaluate the performance of two text-based misconduct detection methods – a mixture distribution model and a Transformer-based text classifier – and apply them to an inference corpus stretching a 42-year time-frame (Sections 4 & 5).

We find that (1) retracted articles have distinctive language patterns and lower lexical diversity than non-retracted articles, (2) paper mill content is detected with an F1 score of 0.93, (3) there is high correlation ($\rho = 0.79$) between the models' estimates and the observed rate of misconduct retractions, and (4) predicted rates of misconduct are hard to interpret but accurately capture trends. More details on the results can be found in Section 6 and a relation to previous work in Section 7.

The **main contributions** of our work are (1) a balanced dataset of retracted and non-retracted articles, containing abstracts and full-text sections, (2) a quantification model based on a mixture distribution to directly estimate the rate of papers containing misconduct from a collection of articles and (3) three Transformer-based classifiers each classifying one of the labels *paper mill*, *randomly generated content* and *falsification*.

We make our code, data, and trained models available on GitHub[1].

2 A Dataset of Retractions

Figure 1 gives an overview of the provenance and size of the datasets used in this study. We utilize a dataset of retracted papers originating from the blog `retractionwatch.com` merged with information from the scientific publication repository `openalex.org` (see Subsection 2.1). Then, we crawl open-access PDF articles from the web to add full-text data to this dataset (see Subsection 2.2). For comparative analysis, we create a reference corpus of non-retracted articles with the same temporal and topical distribution (see Subsection 2.3).

2.1 Retraction Watch Text Corpus

Scientific publishers usually issue paper retractions through their platforms in the form of retraction notices. Typically, retraction notices don't contain extensive information about the backgrounds of a retraction and often go unnoticed by the community (Marcus and Oransky, 2014). To combat this, journalists Ivan Oransky and Adam Marcus started their blog `retractionwatch.com`, where they publish retractions alongside the background stories they manage to investigate. This also led to the creation of the Retraction Watch database consisting of 55,520 entries of retracted articles with associated reasons and nature of retraction.

Reason labels. In the data, there are 106 distinct reason labels, and each record can be assigned multiple reasons (3.6 on average). We identify the biggest reasons linked to scientific misconduct that are potentially recognizable from the paper's text as *Paper Mill*, *Falsification/Fabrication of Data*, and *Randomly Generated Data*. Filtering the retractions by these three reasons results in sub-datasets containing 3,605, 3,090, and 1,016 articles, respectively. Whenever we speak of misconduct hereafter, it will refer to papers tagged with one of these three reasons.

OpenAlex data. The Retraction Watch dataset does not include any content-related data. We use the platform `openalex.org` (Priem et al., 2022) to gather the text of abstracts as well as information about authors, publishers, affiliations, and topics of the papers. OpenAlex is a repository of more than 240 million scholarly documents, which mainly consists of data from the Microsoft Academic Graph (Sinha et al., 2015) and Crossref[2], but also combines information from other metadata sources. Merging the Retraction Watch data with OpenAlex reduces the number of articles in the corpus to 30,901, of which 19,472 have a plain text abstract available. In some cases of retractions, instead of the abstract, we find a retraction notice. Considering that we want to find latent signals of retracted articles, we filter out these cases by excluding abstracts containing the substring "retract".

Domains and fields. OpenAlex employs a three-tiered hierarchy for the research area of a paper, with the *domain* at the top, following a *field* and *subfield* categories. These categorizations allow us to conduct our analyses within and across fields.

2.2 Full-Text Extraction

OpenAlex does not publish any full-text content of articles. Instead, we can often retrieve PDF links of open-access papers through the API. We collect these PDF documents where possible. The PDF documents are converted to raw text by a PDF to

[1] `https://github.com/Christof93/language-of-scientific-misconduct.git`

[2] `https://www.crossref.org/`

Section	Number of Articles
Abstract	19,472
Introduction	6,783
Related Work	1,589
Methods	1,301
Result & Discussion	5,177
Conclusion	5,300

Table 1: Count of retracted articles grouped by successfully extracted sections.

markdown converter[3]. To extract content-related text snippets from the full-text, we employ a simple regular expression matching algorithm that detects sections according to a number of section title variants. We determine the section titles by looking at the frequency distribution of all section titles and choosing titles that fulfill two requirements: (1) the section they preface is likely content-related and (2) they occur very frequently. We group the resulting list of section titles into these five categories: introduction, related work, method, results/discussion, and conclusion (see in Table 5 in Appendix A) for the mapping of titles to categories. This approach ensures that we keep the article contents instead of appendices, tables, references, or even meta information such as the retraction notice prepended to many retracted articles. It also allows filtering and analyzing the content by section type. About 50% of the PDF links allow automated retrieval, and we manually download an additional 1,306 documents to bolster the record. Table 1 shows the total articles and sections obtained.

2.3 Reference Corpus

For our analysis, we construct a parallel reference corpus of non-retracted articles sampled from OpenAlex. Since a random sample of research articles might potentially include a small number of soon-to-be-retracted papers we try to reduce noise by extracting only the top cited articles from OpenAlex, assuming that they are less likely to be fraudulent. For each year and field in the retraction corpus, we collect exactly the same number of articles for the reference corpus. We gather the same information for this dataset and download freely accessible PDF documents where possible.

3 Language Characteristics of Retracted Papers

To analyze linguistic differences between retracted and non-retracted papers, we compare log-odds of word and n-gram occurrences (Subsection 3.1) and investigate the significance of differences in word repetitions (Subsection 3.2).

3.1 Characteristic Expressions

We conduct a log-odds analysis, identifying words that are significantly overrepresented in one corpus versus the other. We apply a chi-square independence test to assess whether frequency differences between corpora were statistically significant, only considering tokens and n-grams meeting the significance threshold of $p < 0.05$. According to the analysis, retracted papers overuse certain adverbs and verbs across all domains in the dataset. Illustrative examples can be found when restricting the dataset by domains and fields.

Computer science. For example, in the field of computer science (a subfield of the physical sciences domain) phrases such as *becoming more and more*, *relatively*, and *developed rapidly* had significantly higher log-odds ratios compared to non-retracted papers. Verbs like *analyzes*, *brought*, and *realized* are also disproportionately more common in computer science retractions.

Physical sciences. In the overall physical sciences domain we find a significantly higher frequency of adverbs such as *erefore* (likely an error in PDF conversion of therefore), *gradually*, *comprehensively*, *vigorously*, and *organically* in retracted works. These adverbs are rather vague and unspecific, which might be a reason why they occur less in evidence-based non-retracted papers.

Social science. In the Social Sciences domain, similar patterns emerged. Adverbs like *accurately, vigorously,* and *scientifically* were more frequent in retracted papers, suggesting that authors needlessly overemphasize results. Non-retracted papers in all domains displayed a higher frequency of cautious and precise language. Terms like *i.e.*, *thereof*, *ideally* and *likely* were more common.

3.2 Lexical Diversity

Examining some of the retracted articles, we notice that they often feature repeated use of the same words, sometimes even within the same sentence. To test this assumption, we calculate the

[3] https://pypi.org/project/pymupdf4llm/

	Sentence Level			Document Level		
POS Tag	Ret TTR	Ref TTR	CLES	Ret TTR	Ref TTR	CLES
VERB	0.982	0.987	0.492	0.794	0.873	0.348
ADJ	0.960	0.974	0.480	0.683	0.771	0.368
NOUN	0.924	0.946	0.456	0.571	0.679	0.326
ADV	0.988	0.992	0.497	0.805	0.890	0.377

Table 2: Mean type-token ratio (TTR) for retracted (Ret) and reference (Ref) texts and Common Language Effect Size (CLES) values between them, at sentence and document level.

type-token ratio (TTR) as a measure of lexical diversity (Baayen, 2001) for selected parts of speech (adjective, adverb, noun, and verb) at both the document and sentence levels (see Table 2).

We consider a Mann-Whitney U test (Mann and Whitney, 1947) to establish the significance of differences in TTR between retracted and non-retracted sentences and documents. The Kolmogorov-Smirnov test (Massey, 1951) confirms that TTR distributions in both corpora are not normal, justifying the use of Mann-Whitney. According to the test, differences between retracted and non-retracted type-token ratios are significant for all selected POS tags ($p < 0.001$).

TTR difference effect size. The Common Language Effect Size (CLES) scores (see Table 2) indicate that differences in lexical diversity within sentences are very small between retracted and reference papers (close to 0.5, which would indicate no difference). Per document, the differences are more pronounced. Lower CLES values at the document level, particularly for nouns (0.3258) and adjectives (0.3681), suggest that retracted papers exhibit lower lexical diversity, meaning they rely more on repetitive phrasing or expressions.

4 Identifying Scientific Misconduct

The language analysis results in Section 2 suggest that retracted papers can be identified through statistical methods based on the differing linguistic structure. Building on this, we focus on specifically identifying retractions involving scientific misconduct next. We present two methods to achieve this, a quantification framework based on a mixture distribution (Subsection 4.1) and a classifier fine-tuned on a pre-trained Transformer model (Subsection 4.2).

4.1 Distributional Quantification Framework

Inspired by recent successes in measuring usage of LLM-generated language, we adapt the *distributional LLM quantification* framework from Liang et al. (2024b) to measure the fraction of research articles that contain language typical of scientific misconduct. The Distributional Quantification Framework (DQF) determines the most likely mixture ratio of two probability distributions pre-calculated on the training data. Let $\mathcal{P}$ denote the distribution of the reference text and $\mathcal{Q}$ that of the type of text we want to quantify. $\mathcal{P}(x)$ and $\mathcal{Q}(x)$ will denote the likelihood of text x under $\mathcal{P}$ or $\mathcal{Q}$ respectively. A collection of texts is described by a mixture of these two distributions:

$$\mathcal{D}_\alpha(X) = (1-\alpha)\mathcal{P}(X) + \alpha\mathcal{Q}(X) \qquad (1)$$

where α is the mixture parameter determining the fraction of examples belonging to the text type.

$\mathcal{P}$ and $\mathcal{Q}$ are estimated from training data – in our case a corpus of non-retracted articles $X_\mathcal{P}$ and a collection of retracted scientific articles $X_\mathcal{Q}$ (Section 2).

To estimate $\mathcal{P}$ as $\hat{\mathcal{P}}$, the method relies on occurrence probabilities of the tokens t from both text corpora. The estimated probability $\hat{p}(t)$ of a token in the reference corpus is defined as:

$$\hat{p}(t) = \frac{\sum_{x \in X_\mathcal{P}} \mathbb{1}\{t \in x\}}{|X_\mathcal{P}|} \qquad (2)$$

i.e., the number of texts containing the token divided by the total number of texts in the specific corpus. Analogously for $X_\mathcal{Q}$, $\hat{\mathcal{Q}}$, and $\hat{q}(t)$. The probability of a text x under $\hat{\mathcal{P}}$ is subsequently given by:

$$\hat{\mathcal{P}}(x) = \prod_{t \in x} \hat{p}(t) \times \prod_{t \notin x} (1 - \hat{p}(t)) \qquad (3)$$

and $\hat{\mathcal{Q}}(x)$ can be derived similarly using $\hat{q}(t)$.

Finally, to infer the coefficient α for an unseen collection of texts $\{x_i\}_{i=1}^{n}$,the DQF uses maximum likelihood estimation under the estimated mixture distribution $\hat{\mathcal{D}} = (1-\alpha)\hat{\mathcal{P}}(X_\mathcal{P}) + \alpha\hat{\mathcal{Q}}(X_\mathcal{Q})$:

$$\hat{\alpha} = \underset{\alpha \in [0,1]}{\operatorname{argmax}} \sum_{i=1}^{n} log((1-\alpha)\hat{\mathcal{P}}(x_i) + \alpha\hat{\mathcal{Q}}(x_i)) \qquad (4)$$

This step will be used to infer an α estimator representing the fraction of texts in a collection exhibiting the style of $\mathcal{Q}$.

4.2 Transformer-based Classifier

As a comparison to the DQF, we also train Transformer-based classifiers. We adopt the commonly used fine-tuning paradigm and train a randomly initialized classification head on top of a pre-trained Transformer encoder model.[4] Specifically, we use four Transformer encoders pre-trained on general text (BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DeBERTa-v3 (He et al., 2021), ModernBert (Warner et al., 2024)) and three encoders adapted to the scientific domain (SciBERT (Beltagy et al., 2019), SciDeBERTa (Kim et al., 2023), and ClinicalBERT (Wang et al., 2023)). We fine-tune these classifiers with a batch size of 16 and a learning rate of $2 \cdot 10^{-5}$, and select the best model across five epochs based on evaluation set performances. To ensure the reliability of our results, we repeat this process for five different randomly generated seeds and report the average performance.

5 Experiments

We evaluate the DQF and Transformer approaches at the level of the document collection and the individual documents. First, we infer the ratio of papers retracted for misconduct from a collection of articles using the DQF (Section 5.1). Second, we compare both approaches by classifying misconduct on the document level (Section 5.2). Finally, we use both the best-performing Transformer from Section 5.2 and the DQF to quantify misconduct through inference on a collection of randomly sampled research articles to explore temporal trends (Section 5.3).

5.1 Quantification of Misconduct

With the DQF, we can infer the ratio of papers involved in misconduct directly from a collection of articles. To evaluate the performance of the DQF on this task, we test its predicted α on constructed mixtures of retracted and non-retracted text fragments and determine how close the estimate is to the true ratio. We follow these steps:

1. We split the data into 50% training and 50% test examples for both the retracted and reference corpus. We run an experiment for each combination of considered sections (e.g. abstract + introduction) and parts of speech.
2. From the test set, we construct 11 variants with different ratios of retracted and reference examples, going from 0% retracted examples to 100% in increments of 10%.
3. We evaluate if the model predicts the appropriate α ratio by considering the difference between prediction and true ratio. For subsequent experiments, we choose the configuration with the closest α estimate to the true ratio.
4. We repeat steps 1 and 2 using the best model from step 3 on subsets of the data filtered by all combinations of domain, field, and retraction reason.

In the following, we report the results of this evaluation approach for the DQF method.

Relying on verbs and adverbs leads to the best estimate. The results of the configuration search from step 3 for the DQF estimator can be found in Table 3. Generally, including the sections *abstract*, *introduction*, and *conclusion* and the POS-tags *verb* and *adverb* leads to the best results. Only relying on adverbs leads to the closest estimates in social sciences, but the bootstrapping variance is high due to data scarcity. the DQF can estimate the ratio α on the document or the sentence level. Running it on the sentence level increases performance across all domains.

Domain	Sections	POS Tags	Mean Error
Health Sciences	A, I	VERB, ADV	0.075 ± 0.011
Life Sciences	A, I	VERB, ADV	0.081 ± 0.010
Social Sciences	A, I, C	ADV	0.063 ± 0.036
Physical Sciences	I, C	VERB, ADV	0.064 ± 0.013

Table 3: Best-performing mixture model per domain with corresponding setting of document sections and POS Tags. A, I, and C stand for abstract, introduction, and conclusion, respectively. ADV means adverbs.

Quantifying paper mill content works better than falsified data. DQF results for specific misconduct retraction reasons can be found in Figure 2. In the case of paper mill quantification, we observe that the method overestimates the true ratio by maximally 15% if the test data entirely consists of non-retracted sentences and underestimates it by around 11% for completely retracted data. The top subplot in Figure 2 shows that the method does not perform well for the retraction reason of falsification. A possible explanation would be that falsification

[4] See Appendix B.1 for more details.

Model	PM	RGC	F&F
BERT	0.905	0.920	0.770
RoBERTa	0.905	0.904	0.779
DeBERTa-v3	0.916	0.911	0.770
ModernBert	0.914	0.903	0.779
ClinicalBERT	0.895	0.886	0.709
SciBERT	0.911	0.914	**0.797**
SciDeBERTa	**0.926**	**0.934**	**0.797**
DQF	0.854	0.798	0.727

Table 4: Different models' F1 on detecting misconduct types: Paper Mill (PM), Randomly Generated Content (RGC), and Falsification and Fabrication of Data (F&F).

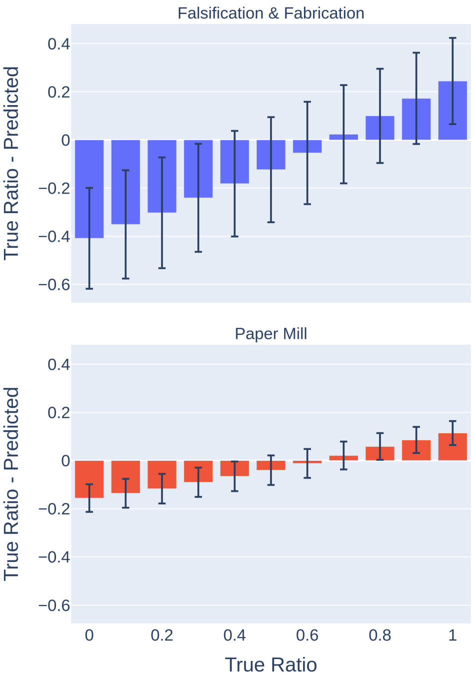


Figure 2: Differences between true ratio and DQF α on a test set of retracted and non-retracted sentences, constructed with ratios $\{0, 0.1, ..., 0.9, 1\}$. Retractions are due to *Falsification* (top) and *Paper Mill* (bottom). Error bars indicate the confidence interval.

happens on the level of experimental results or data and is not directly visible in the article text. This finding is also confirmed by the Transformer-based classifier (see Figure 4b) and inference results (see Figure 8 in Appendix B.3). We also evaluate the performance for the categories *randomly generated content* and *peer review fraud*, which both have similar results (see Figure 5 in Appendix B.2).

5.2 Document-Level Detection

For the misconduct classifier, we look at individual articles. For all misconduct reason subsets of the retraction corpus, we sample an equal-sized subset from the reference corpus matching the distributions of years and scientific fields. Then, we further split the data into training, development, and test sets according to a 60:15:25 split, keeping the label distribution balanced for all sets. We separately fine-tune different Transformer classifiers to perform binary classification for each reason of misconduct. The results are shown in Table 4.

SciDeBERTa is the strongest model on average. Detecting falsification is again the hardest task, as discussed in the previous experiment. Generally, models pre-trained on scientific text perform better than their base models, as we can see when we compare SciBERT to BERT and SciDeBERTa to DeBERTa.

DQF performs worse at classification. To assess the performance of the DQF in a document classification setting and compare it to the Transformer-based classifiers, we replicate the experiment for this approach. We first learn estimators for the distributions $\mathcal{P}$ and $\mathcal{Q}$ on the training set. Then, we infer the estimated α parameter for each document in the development set and measure precision and recall for different classification thresholds. The corresponding precision-recall curve for the paper mill detector on the development set can be found in Figure 6 in Appendix B.2. We take the α threshold producing the best F1 score on the development set to create the final classifier evaluated on the test set. Table 4 shows that the DQF approach performs considerably worse on the test set than the Transformer-based classifiers.

5.3 Misconduct over Time

Next, we turn to estimating scientific misconduct over time by running our best-performing Transformer classifier and the DQF on a much larger inference dataset. We sample 12,000 papers per year from 1980 to 2024 from the OpenAlex API. The sample is divided equally among the four domains defined by OpenAlex: *Health Sciences*, *Life Sciences*, *Physical Sciences*, and *Social Sciences*. Any reportedly retracted articles are excluded from the sampling process. In total, we find 526,876 articles, 390,474 of which have an abstract available. We apply the best DQF model from Section 5.1 for each misconduct reason and each of the four domains to the inference dataset's paper abstracts.

The results for the paper mill model can be found in Figure 3, and for falsification and randomly generated content in Figures 7 and 8 in Appendix B.3.

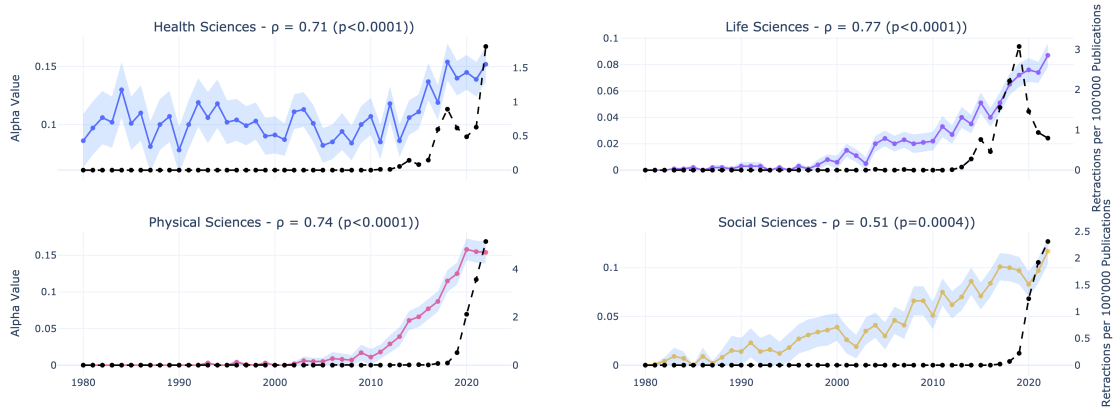


Figure 3: DQF α value signifying the fraction of articles attributed to the *paper mill* category with confidence interval in the shaded area (left y-axis) compared to confirmed paper mill retractions as a fraction of total published articles (right y-axis) per domain. ρ denotes Pearson correlation between the two curves with associated p-value.

To get an impression of the reliability of these results, we overlay the number of confirmed retractions per 100,000 publications listed by the Retraction Watch dataset on the right y-axis. While the α estimate is several orders of magnitude larger than the confirmed retraction ratio, we can observe that the correlation is significant, especially for the domains of Life and Physical Sciences.

Further, we run inference with a Transformer-based classifier on the same dataset. The results in Figure 4 are produced by SciDeBERTa, the best-performing classifier on paper mill and randomly generated content detection. The Figure shows the averaged results across all domains, for the reasons of paper mill and falsification.

Correlation between confirmed and predicted ratios is high. We see a high correlation to the reported retraction for the best-performing paper mill classifier (Figure 4a) and a slightly negative correlation for the falsification classifier (Figure 4b). As mentioned above, we expect that the retractions that fall into the falsification category do not have a strong signal since the falsification might often be limited to study data as opposed to textual content – especially in the health sciences domain where this type of misconduct is most prevalent. The results for the randomly generated content category are omitted from Figure 4, but reported in Table 6 in Appendix B.4. Similar to paper mill, the classification of randomly generated content is significantly correlated with the reported results at $\rho = 0.75$.

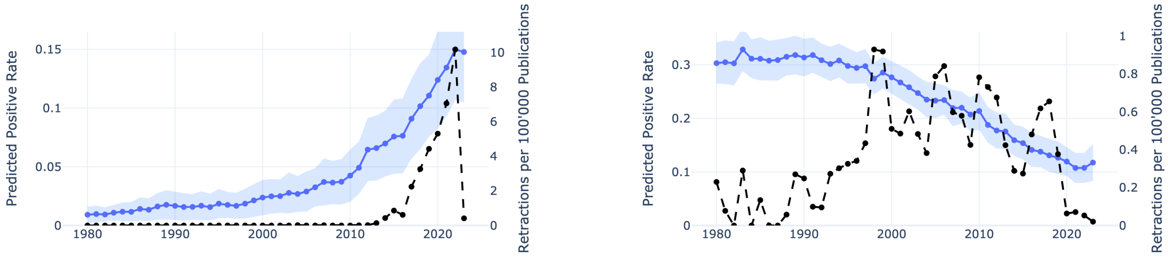
SciDeBERTa classifier estimates are higher than those of other methods. Compared to the DQF results, the SciDeBERTa classifier produces a higher rate of misconduct papers at up to 20% predicted positive rate. This is double the rate predicted by ModernBERT and the DQF. We list the inferred rates of the paper mill papers from the Transformer-based classifier grouped by domains in Figure 9 in Appendix B.4. The estimate seems more stable in the health sciences compared to the DQF results in Figure 3.

6 Discussion

In this section, we revisit the most important results and discuss implications of the findings.

The DQF estimates seem as accurate as the classifier-based ones except in health sciences. We observe that the DQF approach seems to function well in the domains of life sciences and physical sciences and a little less in social sciences. For health sciences, it returns seemingly overestimated and highly varying results. This might be explained by the widespread use of formulaic language in the health sciences domain.

The α is not a precise estimate for the true ratio of misconduct. Our methods estimate that 10–15% of papers are involved in misconduct. The evaluation in Figure 2 shows that the method tends to overestimate the α for a low ground-truth ratio. This indicates that the actual ratio is likely smaller. This method may not be precise enough to reliably detect small effects. Rather than providing an exact fraction of misconduct cases, the α value should be interpreted as the proportion of papers that exhibit writing style similarities with paper-milled papers



(a) $\rho = 0.79$ ($p < 0.0001$)

(b) $\rho = -0.17$ ($p = 0.266$)

Figure 4: SciDeBERTa-based positive prediction rate for detection of *paper mill* (a) and *falsification/fabrication* (b) on the left y-axis compared to confirmed retractions per 100,000 publications on the right y-axis. Pearson correlation ρ between the two curves is given with the associated p-value.

or have a slightly higher-than-average probability of resulting from misconduct. In many cases, the estimated ratio correlates strongly with the reported number of misconduct-related retractions, as the inference study shows.

Trend analysis is a promising application of these models. The practical use case of this becomes clear if we examine the median delay between the date of publication and the date of retraction in our data, which is 475 days, with the 80th percentile at 3.6 years. The DQF method delivers a computationally efficient way to estimate how the incidence of certain types of misconduct changes in real time in a large collection of scientific articles. This could be used, for example, to measure the effectiveness of anti-fraud policies at large publishers.

7 Related Work

Many studies examine the phenomenon of retractions and misconduct (Wray and Andersen, 2018; Candal-Pedreira et al., 2022; Feng et al., 2020; Nath et al., 2006; Parker et al., 2024) but they primarily focus on metadata such as citation information rather than textual content. (Sharma et al., 2024) use the Retraction Watch dataset to analyze author collaboration networks. (Hu and Xu, 2020) and (Vuong, 2019) study linguistic characteristics of retraction notices but not the article content. We do not find any datasets that combine information about retractions with the content of the associated articles. Especially looking at the full text and not only abstracts.

Detection of scientific misconduct. The extensive use of LLM chatbots has led to numerous works focusing on identifying AI-generated content in scientific articles and peer reviews. (Liang et al., 2024b,a) and (Yu et al., 2024) investigate detecting modified sentences by OpenAI's GPT-3 and GPT-4o models. Earlier work from (Gehrmann et al., 2019) trains a model to distinguish human-written sentences from GPT-2 generated ones. (Cabanac and Labbé, 2021) present a detection method that identifies "tortured phrases" which they attribute to using scientific text generators such as SciGen. However, their work is limited to articles from a single journal.

Aside from detecting AI-generated text, some authors explore more general methods to automatically detect fraud and misconduct in science. (Usman and Balke, 2024, 2023) use citation information from retraction cascades to identify potentially retractable articles. (Horton et al., 2020) use Benford's law to identify falsified data specifically. Similar to our work, (Razis et al., 2023) use a Transformer-based model for paper mill content detection. Our work presents a new dataset for this task, which stems from a wider range of science domains and is slightly larger. Further, we extend their analysis by more pre-trained models and an analysis of large-scale inference on a longitudinal dataset.

8 Conclusion

In this work, we find distinct linguistic patterns in articles retracted for misconduct, such as overuse of certain expressions and frequent repetition of adjectives, nouns, and adverbs. Based on these findings, we train a distributional quantification framework and a Transformer-based classifier to track growth trends of scientific misconduct. Our classifier achieves an F1 score of 0.93 in detecting paper mill articles and automatically generated content, making it a viable option for flagging fraudulent papers for human review. Further, we show that a computationally simpler approach based on a mixture distribution model can estimate trends of misconduct in life and physical sciences. However, in health sciences, the Transformer-based classifier performs better. For future work, we will investigate how metadata such as citation networks and affiliation can be incorporated into detecting misconduct and lead to increased performance in cases where the text-based approach does not yield sufficient results.

Limitations

This study has several limitations. First, the underlying Retraction Watch dataset is compiled by volunteer journalists, making its coverage inconsistent. For instance, retractions were reported more frequently during the platforms early years, disproportionately affecting recently published papers, and little is known about the annotation process of reason labels. Additionally, many older retracted papers may no longer be accessible online, as their records have likely been lost.

Furthermore, as was shown in the study, α values do not necessarily reflect the true proportion of paper-milled articles. The estimates can not be taken at face value but serve as a tool to investigate trend evolution.

Finally, availability of text (abstracts and full-text) has limited the size of our text corpus. For the misconduct reason of *falsification*, the dataset is particularly small potentially impacting the accuracy of the resulting classifier. Also, while it might have a positive impact on performance we intentionally exclude non-content-related metadata about publications from the training process to isolate the influence of language. Future work will extend on this.

Ethics Statement

Our work recognizes the ethical implications of predicting misconduct based on individual articles, as false positives could lead to serious reputational harm and may be perceived as slander by the affected authors and/or institutions. Therefore, we emphasize that our method should not be used for definitive individual accusations but rather for statements about collections of articles and trend estimation.

Furthermore, we are aware that releasing an instance-based classification method carries the risk of reverse engineering, allowing malicious actors to manipulate accordingly their writing, in order to evade detection while still perpetrating scientific misconduct. However, we believe that the benefits of transparency outweigh this risk, as security through obscurity is rarely an effective strategy in the long term.

Finally, we acknowledge that our classifier may introduce bias against non-native English speakers, as variations in vocabulary and lexical diversity could influence predictions. Furthermore, low-price text rewriting and translation services may unintentionally produce text that resembles the linguistic patterns associated with misconduct, potentially leading to unfair penalties for individuals. Addressing these biases is a critical area for future work.

Acknowledgements

Special thanks are extended to our colleagues Luca Mazzola and Alexander Denzler at HSLU for their feedback. Furthermore, we thank Oliver Karras and Sören Auer from the Leibniz Information Centre for Science and Technology University Library (TIB) in Hannover for their feedback and support.

References

Anna Abalkina. 2023. Publication and collaboration anomalies in academic papers originating from a paper mill: Evidence from a russia-based paper mill. *Learned Publishing*, 36(4):689702.

R. Harald Baayen. 2001. *Word Frequency Distributions*. Springer Netherlands.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Jeffrey Brainard and Jia You. 2018. What a massive database of retracted papers reveals about science publishing's death penalty'. *Science*. Accessed: 2024-12-23.

Jennifer A Byrne, Yasunori Park, Reese A K Richardson, Pranujan Pathmendra, Mengyi Sun, and Thomas Stoeger. 2022. Protection of the human gene research literature from contract cheating organizations known as research paper mills. *Nucleic Acids Research*, 50(21):1205812070.

Guillaume Cabanac and Cyril Labbé. 2021. Prevalence of nonsensical algorithmically generated papers in the scientific literature. *Journal of the Association for Information Science and Technology*, 72(12):14611476.

Guillaume Cabanac, Cyril Labbé, and Alexander Magazinov. 2021. Tortured phrases: A dubious writing style emerging in science. evidence of critical issues affecting established journals.

Cristina Candal-Pedreira, Joseph S Ross, Alberto Ruano-Ravina, David S Egilman, Esteve Fernández, and Mónica Pérez-Ríos. 2022. Retracted papers originating from paper mills: cross sectional study. *BMJ*, page e071517.

Jana Christopher. 2021. The raw truth about paper mills. *FEBS Letters*, 595(13):17511757.

COPE. 2025. Systematic manipulation of the publishing process: Paper mills. Accessed: 2025-01-14.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniele Fanelli, Julie Wong, and David Moher. 2021. What difference might retractions make? an estimate of the potential epistemic cost of retractions on meta-analyses. *Accountability in Research*, 29(7):442459.

Lingzi Feng, Junpeng Yuan, and Liying Yang. 2020. An observation framework for retracted publications in multiple dimensions. *Scientometrics*, 125(2):14451457.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Joanne Horton, Dhanya Krishna Kumar, and Anthony Wood. 2020. Detecting academic fraud using benford law: The case of professor james hunton. *Research Policy*, 49(8):104084.

Guangwei Hu and Shaoxiong (Brian) Xu. 2020. Agency and responsibility: A linguistic analysis of culpable acts in retraction notices. *Lingua*, 247:102954.

Eunhui Kim, Yuna Jeong, and Myung-Seok Choi. 2023. MediBioDeBERTa: Biomedical language model with continuous learning and intermediate fine-tuning. *IEEE Access*, 11:141036–141044.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024a. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y. Zou. 2024b. Mapping the increasing use of LLMs in scientific papers. In *First Conference on Language Modeling*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):5060.

Adam Marcus and Ivan Oransky. 2014. What studies of retractions tell us. *Journal of Microbiology & Biology Education*, 15(2):151154.

Frank J. Massey. 1951. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):6878.

Sara B Nath, Steven C Marcus, and Benjamin G Druss. 2006. Retractions in the research literature: misconduct or mistakes? *Medical Journal of Australia*, 185(3):152154.

Ivan Oransky, Stephen E Fremes, Paul Kurlansky, and Mario Gaudino. 2021. Retractions in medicine: the tip of the iceberg. *European Heart Journal*, 42(41):42054206.

Lisa Parker, Stephanie Boughton, Lisa Bero, and Jennifer A. Byrne. 2024. Paper mill challenges: past, present, and future. *Journal of Clinical Epidemiology*, 176:111549.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts.

Gerasimos Razis, Konstantinos Anagnostopoulos, Omiros Metaxas, Stefanos-Dimitrios Stefanidis, Hong Zhou, and Ioannis Anagnostopoulos. 2023. Papermill detection in scientific content. In *2023 18th International Workshop on Semantic and Social Media Adaptation &; Personalization (SMAP)18th International Workshop on Semantic and Social Media Adaptation &; Personalization (SMAP 2023)*, page 16. IEEE.

Kiran Sharma, Aanchal Sharma, Jazlyn Jose, Vansh Saini, Raghavraj Sobti, and Ziya Uddin. 2024. Exploring structural dynamics in retracted and non-retracted author's collaboration networks: A quantitative analysis.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (MAS) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, WWW 15, page 243246. ACM.

The Center for Scientific Integrity. 2018. The retraction watch database. ISSN: 2692-4579. [Cited: 2024-08-22].

Muhammad Usman and Wolf-Tilo Balke. 2023. *On Retraction Cascade? Citation Intention Analysis as a Quality Control Mechanism in Digital Libraries*, page 117131. Springer Nature Switzerland.

Muhammad Usman and Wolf-Tilo Balke. 2024. *Tracing the Retraction Cascade: Identifying Non-retracted but Potentially Retractable Articles*, page 109126. Springer Nature Switzerland.

QuanHoang Vuong. 2019. The limitations of retraction notices and the heroic acts of authors who correct the scholarly record: An analysis of retractions of papers published from 1975 to 2019. *Learned Publishing*, 33(2):119130.

Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, et al. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10):2633–2642.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *CoRR*, abs/2412.13663.

K. Brad Wray and Line Edslev Andersen. 2018. Retractions in science. *Scientometrics*, 117(3):20092019.

Sungduk Yu, Man Luo, Avinash Madasu, Vasudev Lal, and Phillip Howard. 2024. Is your paper being reviewed by an LLM? investigating AI text detectability in peer review. In *Neurips Safe Generative AI Workshop 2024*.

A Dataset Creation

This section contains the table with mappings from section category to section title variants that were used to extract full-text of article sections (Table 5).

B Additional Details of the Experiments

In this section, we find all the details needed for reproducing the experiments (Subsection B.1), validation set performance from the DQF evaluation (Subsection B.2), and additional results from the inference study from the DQF (Section B.3 and Transformer-based classifier (Section B.4).

B.1 Used Language Models

All experiments are conducted on a single Nvidia RTX A6000 GPU equipped with 48GB of memory. The models utilized are sourced from the Hugging Face model hub:

- `bert-base-uncased`
- `roberta-base`
- `microsoft/deberta-v3-base`
- `answerdotai/ModernBERT-base`
- `allenai/scibert_scivocab_uncased`
- `KISTI-AI/Scideberta-full`
- `medicalai/ClinicalBERT`

B.2 Additional DQF Evaluation Results

This section contains the DQF evaluation results for peer review fraud and randomly generated content (Figure 5) and the precision-recall curve from finding the DQF detector threshold on the development set (Figure 6).

B.3 Additional DQF Inference Results

In this section, results of the DQF approach on the inference dataset can be found. Figure 7 shows inference of the randomly generated content, and Figure 8 shows that of the falsification estimation model on the large inference corpus.

B.4 Additional Transformer-based Classifier Inference Results

This section contains the remaining inference results of the Transformer-based classifier. Table 6 subsumes the mean positive prediction rate and Pearson correlation for all models and reasons on the inference data. More detailed results per domain and over the years can be found in Figure 9 for paper mill detection by the SciDeBERTa model and for falsification detection by the SciBERT model (best performing on this task) in Figure 10.

Section	Section Title Variants
Introduction	[Objectives, Objective, Background, Introduction]
Related Work	[Related Work, Related Works, State of the Art, Literature Review]
Methods	[Methods, Method, Patients and Methods, Methods and Materials, Methodology]
Result & Discussion	[Discussion, Discussions, Statistical Analysis, Results and Analysis, Results and Discussion, Result and Discussion, Result Analysis, Result, Results, Analysis of Results, Experimental Results, Analysis of Experimental Results, Result Analysis and Discussion, Results and Discussions, Experimental Results and Analysis]
Conclusion	[Conclusion, Conclusions, Authors Conclusions]

Table 5: Mapping of Section Title Variants to Standardized Sections.

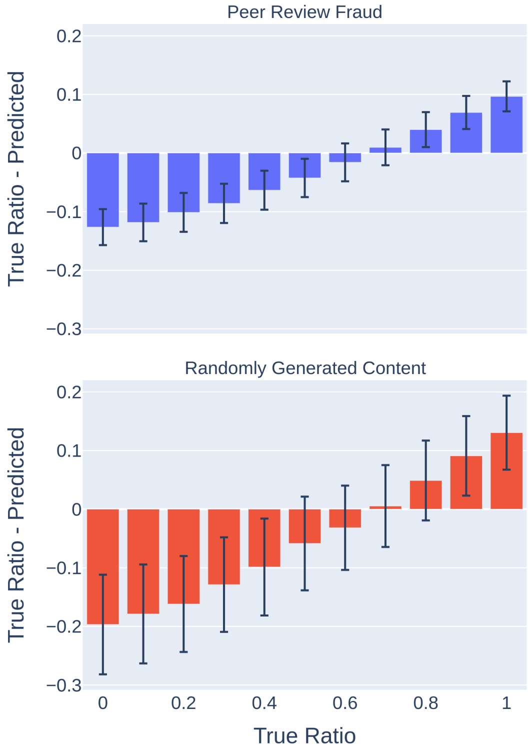


Figure 5: Differences between true ratio and DQF α on a testset partitioned into retracted and non-retracted sentences according to ratios $\{0, 0.1, ..., 0.9, 1\}$. Retractions for reasons of peer review fraud (top) and randomly generated content (bottom).

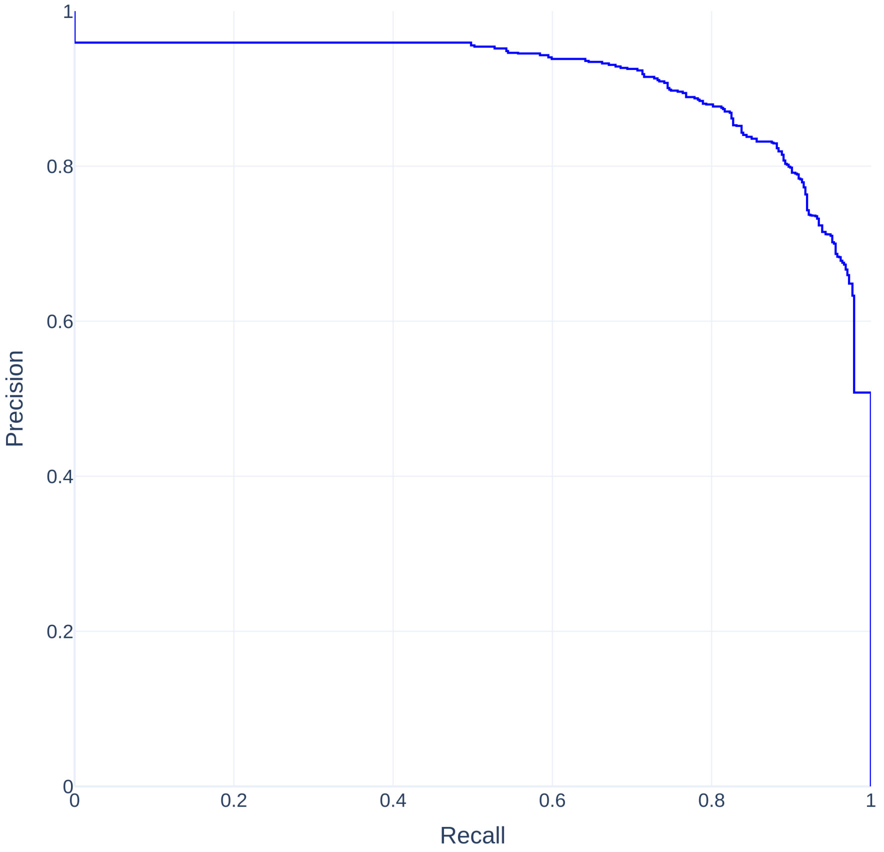


Figure 6: Evaluation of the DQF α used as a detector on the same training and development set as the Transformer-based classifier. The curve shows precision and recall for different thresholds of α values to determine whether a paper should be labeled Paper Mill or not.

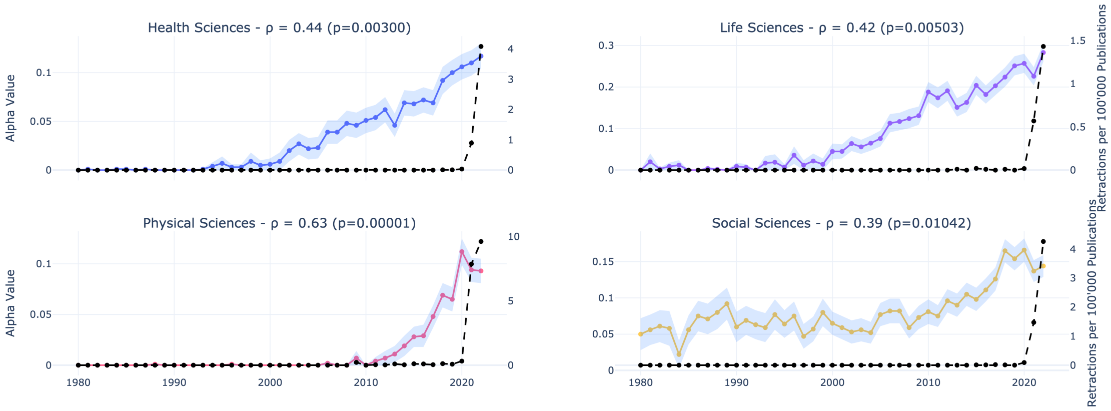


Figure 7: DQF α value signifying the fraction of articles identified as randomly generated content with confidence interval in the shaded area (left y-axis) compared to confirmed randomly generated content retractions (right y-axis) per science domain. ρ denotes Pearson correlation between the two curves with associated p-value.

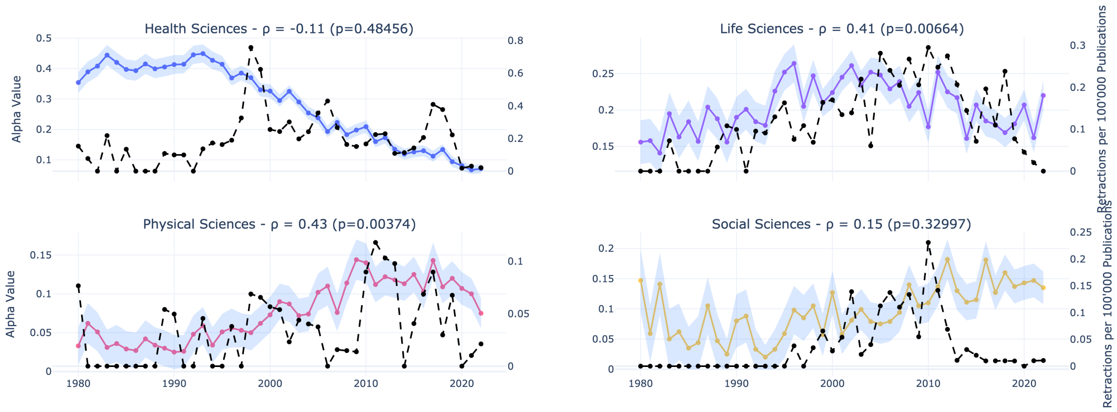


Figure 8: DQF α value signifying the fraction of articles identified as falsification with confidence interval in the shaded area (left y-axis) compared to confirmed falsification retractions (right y-axis) per science domain. ρ denotes Pearson correlation between the two curves with associated p-value.

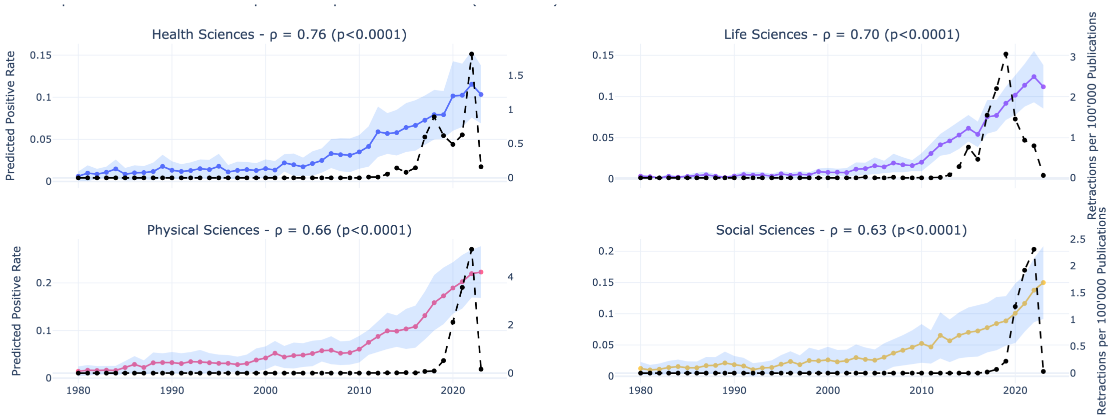


Figure 9: Predicted positive rate of the SciDeBERTa-based classifier on the paper mill detection task (left y-axis) compared to confirmed paper mill retractions (right y-axis) per domain. Pearson correlation ρ between the two curves is displayed with associated p-value.

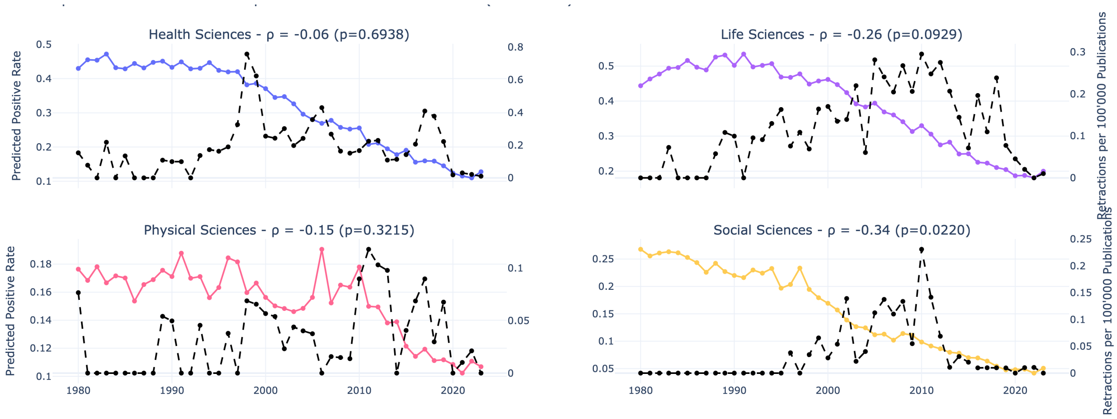


Figure 10: Predicted positive rate of the SciBERT-based classifier on the falsification detection task (left y-axis) compared to confirmed paper mill retractions (right y-axis) per domain. Pearson correlation ρ between the two curves is displayed with associated p-value.

Model	Reason	Mean PPR	Correlation
SciDeBERTa	Paper Mill	0.05	$\rho = 0.78$ (p=0.00000)
SciDeBERTa	Falsification	0.25	$\rho = -0.31$ (p=0.04236)
SciDeBERTa	Generated Content	0.03	$\rho = 0.63$ (p=0.00000)
BERT	Paper Mill	0.07	$\rho = 0.75$ (p=0.00000)
BERT	Falsification	0.26	$\rho = -0.23$ (p=0.12545)
BERT	Generated Content	0.05	$\rho = 0.56$ (p=0.00007)
SciBERT	Paper Mill	0.05	$\rho = 0.78$ (p=0.00000)
SciBERT	Falsification	0.25	$\rho = -0.23$ (p=0.12722)
SciBERT	Generated Content	0.03	$\rho = 0.61$ (p=0.00001)
ClinicalBERT	Paper Mill	0.07	$\rho = 0.76$ (p=0.00000)
ClinicalBERT	Falsification	0.17	$\rho = -0.28$ (p=0.06148)
ClinicalBERT	Generated Content	0.08	$\rho = 0.55$ (p=0.00010)
ModernBert	Paper Mill	0.03	$\rho = 0.79$ (p=0.00000)
ModernBert	Falsification	0.31	$\rho = -0.17$ (p=0.26445)
ModernBert	Generated Content	0.03	$\rho = 0.64$ (p=0.00000)
DeBERTa-v3	Paper Mill	0.05	$\rho = 0.80$ (p=0.00000)
DeBERTa-v3	Falsification	0.39	$\rho = -0.24$ (p=0.11783)
DeBERTa-v3	Generated Content	0.04	$\rho = 0.62$ (p=0.00001)
RoBERTa	Paper Mill	0.03	$\rho = 0.79$ (p=0.00000)
RoBERTa	Falsification	0.26	$\rho = -0.25$ (p=0.10302)
RoBERTa	Generated Content	0.06	$\rho = 0.57$ (p=0.00006)

Table 6: Mean positive prediction rates (PPR) and correlation coefficients for different models and reasons.