Open-source Pipeline for Automated Detection of Unrelated Citations

Anonymous ACL submission

Abstract

Citations are important for ensuring the integrity of scientific literature. However, automated citation verification remains a challenge due to a lack of dedicated datasets and limited research focus. In this paper, we introduce an open-source and automated pipeline that integrates citation retrieval and unrelated citation detection. We have built an annotated dataset to ensure the reliability of our pipeline, which can also be used by others to enhance citation verification tasks. We have also validated the pipeline's applicability to real situations, successfully identifying unrelated citations in real scientific papers. Our work is useful as it assists research integrity scientists to identify potential scientific fraud in a more efficient way.

1 Introduction

Citations are an important part of scientific papers. They are useful for tracking the progression of knowledge, and also for assisting readers in constructing a framework to build new hypotheses (Horbach et al., 2021). During the process of writing scientific papers, however, researchers may make mistakes when citing others, resulting in unreliable citations. These unreliable citations can arise from many reasons: misinterpretation of the cited studies, careless writing, an error in DOI, or other unintentional factors. A notable example is the Vickers case (Vickers, 2017), where an error in the DOI resolution mechanism (da Silva et al., 2023) led to the paper being cited by thousands of completely unrelated studies. However, unrelated citations can also result from deliberate misconduct, such as citations generated by paper mills. Abalkina and Bishop (2023) described paper mills as organizations that sell authorship and citations for publications placed in legitimate journals.The citations generated by papermills are often meaningless and irrelevant to their cited studies. Such unrelated citations can lead to the distortion of citation counts and result in potential wrong decisions when these numbers are used in real life (e.g. for individual promotion or to calculate research impact).

We introduce our automated pipeline for detecting such unrelated citations, the ones that are unrelated and completely irrelevant to their cited studies (An example in Appendix 5). Detecting such citations can be helpful for identifying citation manipulation behaviors and potential paper mills. To our knowledge, many research integrity scientists currently rely on manual methods to collect and analyze academic articles. While some may use AI chatbots (e.g., ChatGPT (Achiam et al., 2023), DeepSeek (Guo et al., 2025)) to assist their work, these tools may have limitations for large-scale analysis due to their high costs and data privacy restrictions. To the best of our knowledge, no existing automated pipeline or systematic method can, given a paper's DOI, verify whether its citations contextually correspond to the content of the cited works.

To address this gap, our open-source pipeline integrates three key functions: citation extraction, cited article retrieval, and textual similarity assessment. Given the DOI of a paper, the pipeline automatically verifies citations by comparing each citation context with the corresponding cited abstract (when accessible). To validate the effectiveness of our pipeline, we have also collected and annotated different datasets to check the performance of methods integrated in our pipeline. These datasets will also be open-sourced and can be used by others for different purposes. We also applied our pipeline on random DOIs that have cited retracted papers to spot potential unrelated citations. Our pipeline[1] enables research integrity scientists to efficiently screen citations across scientific papers more effi-

[1] Link will be provided upon acceptance

ciently, making it easier and faster to detect and combat unrelated citations.

2 Related Work

Citations in scientific papers have been a research subject in natural language processing (NLP) and scientometrics, ranging from identifying citation intent/function and sentiment to citation recommendation.

Currently, there are tools such as Nicholson et al. (2021) that identify citation contexts and their functions in research papers to showcase how a certain work is cited throughout the literature. Other researchers in NLP showed efforts in providing annotated datasets for citation function/sentiment (Athar, 2011) and approaches for automatic classification tasks (Teufel et al., 2006). Some examples to the labels used in classification tasks in such works are *"negative", "positive", and "objective"* (Liu, 2017), or distinguishing *"critical"* from *"non-critical"* citations (Te et al., 2022).

Following such use of citations in the NLP literature, citation recommendation systems have emerged as an important application. It can be viewed as the inverse task of detecting unrelated or inappropriate citations. For example, Buscaldi et al. (2024) frame citation prediction as both a Mask-Filling and a Named Entity Recognition problem, proposing transformer-based models enhanced with NLP heuristics.

To the best of our knowledge, the only prior study that directly tackles the automatic detection of unreliable citations is Sarol et al. (2024), who assess citation integrity in biomedical literature, which is more similar to our task. For this purpose a total of 3,063 citations are annotated and labeled in 8 different classes: *Accurate, Contradict, Not_Substantiate, Irrelevant, Oversimplify, Misquote, Indirect* and *Etiquette*. The proposed approach is structured into 2 steps: (1) extracting relevant evidence from the referenced paper, and (2) predicting a label by integrating the citation context with the retrieved evidence. In this particular setup, both steps are reported to be challenging for both human and machine. Despite human training and providing guidelines, the consistency of annotators remained lower than expected. Automatic annotation with these 8 labels was also found to be extremely challenging for the models they tested (e.g. fine-tuned BERT model, GPT-3.5-turbo, GPT-4). Therefore the task was redefined using only three labels: *ACCURATE* that groups the two labels *Accurate* and *Indirect*. *NOT_ACCURATE* by grouping the *Contradict, Not_Substantiate, Oversimplify, Misquote and Etiquette* labels. *IRRELEVANT* that is composed of the sole previous *Irrelevant* label. The main difference between this work and ours is our accessible pipeline and our inclusion of unrelated citations, which is completely not in the same research field with the cited article.

In this work, we focus on detecting unreliable citations, a task that requires retrieving, building datasets, integrating information from the cited paper and textual similarity assessment. While Liu et al. (2024) addressed only the feasability of automatic detection of off-topic citations, our approach extends a pipeline by proposing a structured framework to systematically evaluate citation relativeness within its broader context.

3 Pipeline Architecture

Our pipeline integrates three core functions: (1) citation context extraction, (2) cited abstract retrieval, and (3) textual similarity assessment (Appendix A). Given a list of DOIs, the pipeline first retrieves full-text XML files from the PubMed Central (PMC) database (Roberts, 2001). For each XML file, citation contexts are extracted, and the corresponding cited abstracts are fetched via PubMed and CrossRef APIs. The reliability of each citation is then assessed by calculating textual similarity between the citation context and the cited abstract to detect unrelated citations. Due to copyright restrictions, not all cited articles are fully accessible, but our pipeline can easily be extended to query other data sources.

To evaluate the retrieval coverage of our pipeline, we used a sample of 2000 papers that cite retracted publications. We retrieved[2] this data from the Problematic Paper Screener's (PPS)(Cabanac et al., 2022) *Feet of Clay Detector* which automatically flags publications that cite retracted works for post-publication reassessment. Our experiments revealed that approximately 30% of DOIs are accessible as full-text documents, and 50% of cited abstracts are retrievable with CrossRef[3] and PubMed APIs [4] (Table 1).

[2]Downloaded on 15th of April 2025
[3]https://api.crossref.org/swagger-ui/index.html
[4]https://www.ncbi.nlm.nih.gov/home/develop/api/

Total Random DOIs	Retrievable DOIs	Abstracts of Retrievable DOIs	Retrievable Abstracts
30	10	1297	631
50	15	2266	1007
50	11	1221	717
50	14	2247	1167
100	31	5254	2848

Table 1: Number of Retrievable Full-Text DOIs within Random DOIs Sampled from *Feet of Clay* Dataset

3.1 Citation Context Extraction

We define the citation context as the sentence containing the reference marker (e.g., citation number or author-year format). To ensure sufficient context for analysis, we expand this to include both the preceding and following phrases when the original citation context is shorter than 125 characters. This extended context provides more meaningful text for textual similarity assessment.

3.2 Cited Article Retrieval

For each extracted citation context, the pipeline first identifies the referenced articles by prioritizing their PubMed PMIDs [5]. If a PMID is unavailable, it extracts the article's DOI instead. After that, the pipeline queries the PubMed and CrossRef APIs to locate these cited articles and retrieves their abstracts for further analysis.

3.3 Textual Similarity Assessment

We have experimented with two approaches: pretrained language models (PLMs) approach and text-overlapping approach to assess the textual similarity between citation context and the cited abstract. To perform these experiments we build a synthesized dataset that that contains both related and unrelated citations 3.3.1. We then pick the best method to integrate into our pipeline to check its efficacy in real world situations.

3.3.1 Experimental Dataset

Our dataset includes two types of citations:

Related Citations: Citations that are relevant and correspond to the cited works. These citations were collected from trusted journals such as *The Lancet*, *Cell*, and *Joule* using Elsevier[6] and CrossRef APIs. These journals are selected for their rigorous peer review, assumpting thus citation reliability.

Unrelated Citations: Citations belonging to another research topic and totally irrelevant to the cited paper. These were created artificially by pairing citation contexts from our related citations with irrelevant abstracts. To find these abstracts, we searched CrossRef API using five unrelated keywords, collecting about 100 abstracts per keyword (500 total). Each citation context is then randomly matched with 15 abstracts from three different keywords. This mimics how unrelated citations often appear in practice.

3.3.2 PLMs Approach

This approach calculates cosine similarity scores between embeddings of citation contexts and their corresponding cited abstracts. Embeddings are generated using six models: BERT (Devlin et al., 2018), SBERT (Reimers and Gurevych, 2019), DistilBERT (Sanh et al., 2019), and T5 (Raffel et al., 2020). For each model, we extract embeddings from the final hidden state, mask padding tokens, and average the embeddings of all tokens in the input text.

3.3.3 Text-Overlapping Approach

This approach calculates textual similarity based on overlapping text proportions. It is faster and has lower computational demands compared to PLMs. The workflow involves two steps: (1) Remove stopwords (using NLTK (Bird et al., 2009)) from both the citation context and the cited abstract. (2) Apply different metrics: *BLEU*, *ROUGE*, and *Jaccard* to calculate the textual similarity between citation context and the corresponding cited abstract.

3.3.4 Threshold Selection

For every method in both approaches, we use ROC curve to determine the optimal similarity score threshold, calculated within our experimental dataset. If the similarity score between citation context and the cited abstract is higher than the threshold, then the citation is considered related, vice-versa.

Table 2 represents the performance of different methods on our dataset.

[5]Unique identifiers for PubMed papers

[6]https://dev.elsevier.com/

Method	Experimental Dataset		
	F1 Score	**Precision**	**Recall**
SBERT	0.98	0.99	0.98
DistilBERT	0.90	0.90	0.90
BERT	0.90	0.91	0.89
T5	0.70	0.73	0.68
Rouge	0.76	0.72	0.79
Bleu	0.68	0.71	0.65
Jaccard	0.86	0.93	0.81

Table 2: Performance of citation verification methods on our synthesized experimental dataset

Method	Annotated Dataset (Test-set)		
	F1 Score	**Precision**	**Recall**
SBERT	0.88	0.80	0.98
DistilBERT	0.77	0.66	0.93
BERT	0.75	0.62	0.94
T5	0.52	0.45	0.63
Rouge	0.69	0.63	0.76
Bleu	0.66	0.66	0.65
Jaccard	0.68	0.54	0.92

Table 3: Performance of citation verification methods on annotated dataset

4 Application of the pipeline in real situations

We did two experiments to analyze the possibility of applying the pipeline in real situations. (1) Test on *Annotated Dataset* (Test-set): We built an annotated dataset which includes both related and unrelated citations extracted from real papers to test each method integrated in our pipeline 4.1. (2) *Simulated Deployment:* We run our pipeline through 150 random DOIs in the *Feet of Clay dataset* (Cabanac et al., 2022), simulating how research integrity scientists might use it to flag potential unrelated citations in practice 4.2.

4.1 Test on *Annotated Dataset* (Test-set)

Dataset: We manually collected both related and unrelated citations in scientific papers across different research fields. We have 430 unrelated and 113 related citations in this dataset. Most of the unrelated citations are extracted from papers in Vickers's case. Two annotators independently labeled each citation by comparing its context with the corresponding cited paper's abstract. Citations were retained in the dataset only upon full inter-annotator agreement, with 24 citations excluded.

Test: For each citation in the annotated dataset, we apply different methods from both approaches. The pipeline classifies citations as reliable if their similarity score exceeds the threshold, and unreliable otherwise. The classification performance is calculated using standard metrics: F1-score, precision, and recall. The results are in the table 3. We noticed a drop of performance in the *Annotated Dataset*, this is mainly due to the unbalanced data, and the optimal threshold determined only on *Experimental Dataset*. We can see that SBERT has the best F1 score, so we choose to use this method for the following simulated deployment test.

4.2 Simulated Deployment

We sampled 150 DOIs randomly from the *Feet of Clay* dataset and ran our pipeline using SBERT as its textual similarity verification method. There are, in total, 40 retrievable DOIs from PMC among the 150 DOIs, and in total 2891 retrievable cited abstracts among 5734 cited abstracts. After eliminating 18 citations with incomplete citation context or cited abstract, 38 citations within 13 DOIs have been marked as *Unrelated* by our pipeline. Among these 38 citations, two annotators verified manually and agreed on spotting 15 *Unrelated* citations in 6 different DOIs. 22 of the 38 citations were annotated as *Not sure*, and only 1 citation was annotated as *Related*. This demonstrates the pipeline's potential to assist research integrity scientists in efficiently detecting potential unrelated citations, even with partial data accessibility.

5 Conclusion and Future Work

In this article, we presented an open-source pipeline for automated citation relativeness verification. We tested the pipeline on annotated dataset collected from scientific papers, and it also successfully spotted potentially unrelated citations in a random sample of publications. Our work can be useful for assisting publishers, conference commitees or research integrity scientists to spot potential unrelated citations easier and we look forward to its future use. Future work will focus on: (1) enhancing the retrieval coverage of our pipeline through integration with other open-source APIs. (2) developing computationally lightweight methods to improve the cost-efficiency of citation verification. (3) establishing a systematic taxonomy to categorize nuanced types of citation reliability.

Limitations

Our pipeline uses the abstracts of cited papers for verification, operating under the assumption that they contain sufficient information to assess citation validity. While this approach proves effective for identifying unrelated citations, it faces challenges in detecting nuanced discrepancies such as subtle misrepresentations or partial inaccuracies. Similarly, our textual similarity-based methods faces the same limitations in addressing such complex cases.

The application of our pipeline in real situations is still limited, as our annotated dataset is not very large and our simulated deployment is only on the *Feet of Clay* dataset. The *Feet of Clay* dataset focuses on papers citing retracted articles, which may not represent all types of unreliable citations we can come across in the literature.

A further constraint arises from copyright restrictions: our pipeline can only verify citations to open-access papers, resulting in lower retrieval rates and a reliance on the availability of publicly accessible content.

Acknowledgments

Will be added upon acceptance.

References

Anna Abalkina and Dorothy Bishop. 2023. Paper mills: a novel form of publishing malpractice affecting psychology. *Meta-Psychology*, 7.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Awais Athar. 2011. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*, pages 81–87, Portland, OR, USA. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*.

Davide Buscaldi, Danilo Dessí, Enrico Motta, Marco Murgia, Francesco Osborne, and Diego Reforgiato Recupero. 2024. Citation prediction by leveraging transformers and natural language processing heuristics. *Information Processing & Management*, 61(1):103583.

Guillaume Cabanac, Cyril Labbé, and Alexander Magazinov. 2022. The 'Problematic Paper Screener' automatically selects suspect publications for post-publication (re)assessment. *Preprint*, arXiv:2210.04895.

Jaime A. Teixeira da Silva, Neil J. Vickers, and Serhii Nazarovets. 2023. From citation metrics to citation ethics: Critical examination of a highly-cited 2017 moth pheromone paper. *Scientometrics*, 129(1):693–703.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Serge Horbach, Kaare Aagaard, and Jesper W. Schneider. 2021. Meta-Research: How problematic citing practices distort science. MetaArXiv aqyhg, Center for Open Science.

Haixia Liu. 2017. Sentiment analysis of citations using word2vec. *CoRR*, abs/1704.00177.

Qinyue Liu, Amira Barhoumi, and Cyril Labbé. 2024. Miscitations in scientific papers: dataset and detection. A workshop paper for The Bibliometric-enhanced Information Retrieval workshop series (BIR 2024).

Josh M. Nicholson, Milo Mordaunt, Patrice Lopez, Ashish Uppala, Domenic Rosati, Neves P. Rodrigues, Peter Grabitz, and Sean C. Rife. 2021. scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative Science Studies*, 2(3):882–898.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Richard J Roberts. 2001. Pubmed central: The genbank of the published literature.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Maria Janina Sarol, Shufan Ming, Shruthan Radhakrishna, Jodi Schneider, and Halil Kilicoglu. 2024. Assessing citation integrity in biomedical publications: corpus annotation and nlp models. *Bioinformatics*, 40(7):btae420.

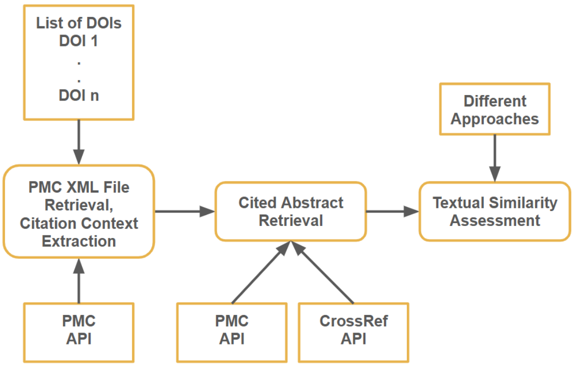

Figure 1: Structure of our pipeline

Sonita Te, Amira Barhoumi, Martin Lentschat, Frédérique Bordignon, Cyril Labbé, and François Portet. 2022. *Citation Context Classification: Critical vs Non-critical*. Association for Computational Linguistics, Gyeongju, Republic of Korea.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110, Sydney, Australia. Association for Computational Linguistics.

Neil J. Vickers. 2017. Animal communication: When i'm calling you, will you answer too? *Current Biology*, 27(14):R713–R715.

A Appendix

N°	Label	Citation context	Abstract of cited paper
(1)	Related	Differently, transformer is a type of neural network mainly based on self-attention mechanism [35], which can provide the relationships between different features.	The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. ...
(2)	Unrelated	The contribution of organic materials has been well acknowledged in the application of electronic devices [28–32].	Male moths compete to arrive first at a female releasing pheromone. A new study reveals that additional pheromone cues released only by younger females may prompt males to avoid them in favor of older but more fecund females.

Table 4: Example of related and unrelated citations

Similarity Score SBERT	Citation Context	Abstract of Cited Paper
0.02	MicroRNAs are highly sensitive to environmental stressors, as is well demonstrated in the lung for cigarette smoke [297] and airborne pollutants [298].	The liming/unhairing operation is among the important processes of the leather industry. It generates large amounts of effluent that are highly loaded with organic hazard wastes. Such effluent is considered one of the most obnoxious materials in the leather industry, causing serious environmental pollution and health risks. The effluent is characterized by high concentrations of the pollution parameters. Conventional chemical and/or biological treatment of such wastewater is inefficient to meet the required limits of standard specifications, due to the presence of resistant and toxic compounds. The present investigation deals with an effective treatment approach for the lime/unhair effluent using the Fenton reaction followed by membrane filtration. The experiment was extended to a laboratory pilot-scale in a continuous treatment study. In this study the raw wastewater was treated with the predetermined Fenton's optimum dose followed by membrane filtration. The wastewater was efficiently treated and the final effluent met the standards for unrestricted water reuse.

Table 5: A potential unrelated citation from *Simulated Deployment* (Also spotted in PubPeer:)