Scientometrics
https://doi.org/10.1007/s11192-025-05524-7

Presenting a classifier to improve the identification of research journal publications in OpenAlex

Nick Haupka[1]

Received: 6 August 2025 / Accepted: 15 December 2025

Abstract
This paper introduces a document type classifier with the purpose to optimise the distinction between research and non-research journal publications in OpenAlex. Based on open metadata, the classifier can identify non-research or editorial content within a set of classified articles and reviews (e.g. *paratext*, *abstracts*, *editorials*, *letters*), which for example is relevant for bibliometric studies, university rankings and academic procedures. In this respect, OpenAlex shows issues in classifying journal research contributions, tending to overestimate them compared to other databases, due to the promotion of non-research items to research items. The classifier presented in this study achieves an F1-score of 0.95, indicating a potential improvement in the data quality of bibliometric research in OpenAlex when applying the classifier on real data. In total, 4,589,967 out of 42,701,863 articles and reviews could be reclassified as non-research contributions by the classifier, representing a share of 10.75%.

Keywords OpenAlex · Document types · Machine learning · Classification · Bibliometrics

Introduction

Scholarly databases like OpenAlex (Priem et al., 2022), Crossref (Hendricks et al., 2020) and Semantic Scholar (Kinney et al., 2023) index millions of research publications each year. In this process, the document type plays a decisive role in distinguishing different types of works such as *editorials*, *case reports*, *journal articles* and *reviews*. The decisions made by database providers to classify publications into document types have a profound impact on numerous bibliometric procedures and performance evaluations. For example, university rankings like the Leiden Ranking Open Edition (Van Eck & Waltman, 2024) restrict their performance analysis on certain document types such as *articles* and *reviews*, while excluding document types like *editorials* and *letters*. Another measurement that relies on document types given by scholarly databases is the Journal Impact Factor. The Journal Impact Factor represents the influence of a journal by counting citations of *articles*

✉ Nick Haupka
nick.haupka@sub.uni-goettingen.de

[1] Göttingen State and University Library, University of Göttingen, Platz der Göttinger Sieben 1, 37073 Göttingen, Germany

and *reviews* published in a journal.[1] Document types are also frequently used as a restriction criterion for research publications, for example when analysing the dissemination of open access in scholarly communication (Piwowar et al., 2018) or when examining the influence of transformative agreements on hybrid journals (Jahn, 2025b).

However, the typology to describe works in bibliometric databases as well as the methods used to assign document types vary between data sources and providers. A study by Haupka et al. (2025) shows that OpenAlex tends to overestimate the assignment of the document type *article* when comparing it to Scopus, Web of Science, Semantic Scholar and PubMed. Alperin et al. (2024) made a similar observation, noting deviations in the classification of document types between OpenAlex and Scopus. Studies also show that the methods for applying document type classification by scholarly data providers are not always precise as the document type assigned within a database can sometimes differ from the actual document type. For instance, Donner (2017) investigated a sample of document types retrieved from Scopus and Web of Science and compared them to the official publisher websites and found discrepancies in more than 17% of the cases. Sometimes, document types are also not covered in bibliographic databases. For example, an analysis of data journals by Jiao et al. (2023) has shown that OpenAlex lacks the classification of data papers which are assigned to an independent document type in Scopus and Web of Science but considered as journal articles in OpenAlex. The same applies to case reports, which are considered journal articles in Scopus and OpenAlex but are regarded as a separate document type in Semantic Scholar and PubMed (Haupka et al., 2025). Delgado-Quirós and Ortega (2024) compared typologies of document types of eight different bibliometric databases, including the databases OpenAlex, Crossref, The Lens, Scilit and Dimensions and found a high degree of consistency between document types retrieved from these databases, which can be attributed to Crossref as the main source of all databases mentioned.

This study focuses on the OpenAlex service, which is a promising (new) data source for bibliometric analyses as it provides its data free of charge and, in some aspects, shows comparable quality to established databases like Scopus and Web of Science (Culbert et al., 2025; Jahn, 2025a). However, as of 2023, OpenAlex has lacked a clear distinction between research articles and reviews. Furthermore, over 99% of journal items have been classified as articles, which seems quite high compared to other databases (Haupka et al., 2025). In 2024, the team behind OpenAlex, OurResearch, has therefore announced and implemented several changes to its document type classification.[2] First and foremost, it included the classification of preprints and reviews. Further it integrated document types from PubMed to relabel a set of articles as editorials and letters. Also, it extended their list-based approach for detecting paratexts and editorial material. The effects of this changes were documented by Haupka et al. (2024). While the implementation of document types from PubMed reduces the amount of works with the type *article*, a gap between the document type classification of OpenAlex and proprietary databases like Scopus and Web of Science still remains.

In recent years, the application of machine learning algorithms have become increasingly popular in open science and especially in bibliometric databases (Jeangirard, 2022). A prominent example is Semantic Scholar which uses machine learning classifiers for,

[1] https://support.clarivate.com/ScientificandAcademicResearch/s/article/Journal-Citation-Reports-Document-Types-Included-in-the-Impact-Factor-Calculation?language=en_US

[2] https://github.com/ourresearch/openalex-guts/blob/main/files-for-datadumps/standard-format/RELEASE_NOTES.txt

among others, author disambiguation and field of study classification (Kinney et al., 2023). Another example is OpenAlex which uses a probability model to detect the language of a work. Further it uses machine learning classifiers to match records with topics and SDGs.[3] Also, the French Open Science Monitor uses machine learning algorithms to detect datasets and software (Bracco et al., 2025).

In line with these developments, this article presents a machine learning classifier that helps to correct OpenAlex's estimates of research documents. The classifier filters (presumed) research contributions in OpenAlex based on open metadata such as author count and number of references[4]. Similar efforts were made by Charalampous and Knoth (2017) who created a classifier that takes metadata into account to distinguish between papers, theses and slides in (open access) repositories. Metadata for their classifier were extracted and aggregated from PDFs. They consisted of the number of authors, the total number of words in a document, the number of pages and the average number of words per page. Overall, this approach resulted in a high F1-score (0.96). The F1-score describes the harmonic mean between precision and recall. The F1-score can take values between 0 and 1, whereby a high F1-score can be considered as a good classification result (see also the Appendix). A similar approach was conducted by Caragea et al. (2016) who extracted file and text specific features from documents to differentiate between books, slides, theses and papers using machine learning techniques. A more traditional approach to address the issue of misclassification of journal articles was proposed by Maisano et al. (2025). It leverages discrepancies between pairs of databases (e.g., Scopus and Web of Science) to automatically flag potentially misclassified articles. A subsequent manual analysis focusing on this subset of "suspect" items then reveals actual errors and attributes them to the responsible database. Last but not least, the Collaborative Metadata Enrichment Taskforce (COMET) has also developed a resource type classifier based on a large language model that can differentiate DataCite records into classes such as books, conference papers, software and journal articles.[5]

The remainder of this article is organised as follows. Section "Methodology" provides the methodological part, describing the data sources used and the construction of the classifier, including a brief explanation of the selected features. Section "Results" contains the results achieved with the classifier. The results are then compared in Section "Validation" with real data from OpenAlex and Scopus, including a small manual sample analysis. And finally, the main findings of this article are presented and discussed in Section "Discussion and conclusion". The "Appendix" section provides additional tables and equations that are useful for this study.

Methodology

Data

Creating a gold standard dataset for validating document types is difficult. Data sources like Crossref, OpenAlex and PubMed mostly rely on publisher information when classifying

[3] https://docs.openalex.org/api-entities/works/work-object

[4] Some results of the classifier were previously briefly documented in a blog post in October 2024 (Haupka, 2024). Since then, minor changes were applied to the classifier.

[5] https://huggingface.co/cometadata/generic-resource-type-lora-qwen2.5-7b

document types[6,7] As Haupka et al. (2025) have demonstrated, this can lead to inconsistencies in databases because each publisher can decide independently how to classify a publication. Also, data sources like Crossref and PubMed apply different taxonomies which makes it difficult to compare document types from different data sources.

For this analysis, it was decided to use PubMed's document types taxonomy, as previous analyses have shown that the data source reaches comparable accuracy against Scopus and Web of Science in regard to the classification of research articles and reviews (Haupka et al., 2025). PubMed, developed and maintained by the National Center for Biotechnology Information and the U.S. National Library of Medicine, is a free online service for searching biomedical and life science literature.[8] It has its own document type classification system, which involves curation and review by professionals.[9]

In order to validate the classifier's results on other data sources like Semantic Scholar or Scopus in a later phase, a shared corpus was constructed between OpenAlex, Scopus, Web of Science, Semantic Scholar and PubMed with limitation to the publication type journal and the publication years 2012 to 2022. Note that the publication type describes the source of a text (e.g. a journal), while the document type indicates the genre of text (e.g. a review) (Haupka et al., 2024).

Document types from PubMed were reassigned to one of two classes: research and non-research. Document types like *article*, *review* and *clinical study* were considered research, while document types such as *retraction note*, *editorial* and *news* were considered non-research, as they lack academic discourse. A complete mapping table can be found in the code supplement[10] or in the Appendix (see Table 4). Note that case reports were originally categorised as research. However, it is debatable whether a case report can be considered a research contribution in this context, as a case report is usually very short compared to typical research articles and reviews. The distinction between research and non-research texts is also hampered by predatory publishing and paper mills, which aim to imitate scientific behaviour. A specific exclusion of journal titles was not taken into account in this study.

Analysis data was partially retrieved from the German Competence Network for Bibliometrics (Schmidt et al., 2025). This includes data from Scopus (July 2024 snapshot) and OpenAlex (August 2024 snapshot). Besides data from OpenAlex, Crossref and PubMed was used to train the classifier. The results of the classifier, including scores, are made available on Zenodo (Haupka et al., 2025).

As of August 2024, OpenAlex includes 17 types of documents, when restricting to the primary location source type *journal* and publication years 2014 to 2023. Research articles and reviews are the most prominent types with overall 93.9%. In all cases, a document type was linked to a document. Note that 8,445,541 (11.27%) of works do not have a source type in OpenAlex (for the publication years 2014 to 2023).

[6] https://crossref.gitlab.io/knowledge_base/docs/topics/content-types/
[7] https://pubmed.ncbi.nlm.nih.gov/help/
[8] https://pubmed.ncbi.nlm.nih.gov/about/
[9] https://www.nlm.nih.gov/bsd/indexfaq.html
[10] https://github.com/naustica/openalex_doctypes/blob/classifier/queries/oal_doc_dataset_extended.sql

Approach

Following (Charalampous & Knoth, 2017), the assumption was made that document types have specific metadata characteristics which can be used to differentiate distinct types of works. This premise also relies on the analysis of Haupka et al. (2025) who demonstrated that works that are categorised as research contributions (e.g. reviews and articles) have, among other metadata, different average numbers of authors and references as non-research contributions (e.g. editorials, paratexts and letters).

The experiments resulted in ten features (F1–F10) that were chosen to train the classifier (see Table 1). Some of these features can be taken from the databases (Crossref and OpenAlex) without further problems while others have to be computed. This applies to F2, F3 and F4.

It was decided to integrate and train the classifier on some bibliographic data from Crossref instead of OpenAlex, as Crossref covers certain metadata fields more comprehensively, especially when looking at the provision of page numbers (Delgado-Quirós & Ortega, 2024). The page number was calculated by computing the absolute number of the difference of the first page and the last page. In cases where the first or last page number was missing it was decided to give the document a page count of 1. The idea behind this choice was that every publication must have a minimum of one page. Of course, it would also be possible to fill in missing values differently (e.g. filling all incomplete values with a value that indicates a missing value), which would eventually lead to different results. About 15% of journal items in the dataset contain no metadata about page information. Note that regular expressions were applied on the page strings to obtain better results when calculating the page numbers. An additional filter was applied in Section "Appendix", namely: If the journal issue contains the words "%sup%" (for supplementary materials) or "%meet%" (for meeting abstracts) the publication was considered as non-research. This affects about 803,688 (1.88%) articles and reviews in the dataset. To verify the effectiveness of this decision, a sample of 50 publication was drawn, which showed that each publication was indeed not a research journal article but rather an abstract or supplementary material.

Table 1 Selected features for the classifier

Feature	Type	Retrieved from	Manually calculated
F1: has abstract?	Boolean	Crossref	False
F2: title word count	Integer	Crossref	True
F3: page count	Integer	Crossref	True
F4: author count	Integer	Crossref	True
F5: has license?	Boolean	Crossref	False
F6: number of citations	Integer	Crossref	False
F7: number of references	Integer	Crossref	False
F8: has funding information?	Boolean	Crossref	False
F9: number of affiliations	Integer	OpenAlex	False
F10: has OA url?	Boolean	OpenAlex	False

Table 2 Test and validation set results

Algorithm						
	Measure	LR	RF	KNN	AdaBoost	Baseline
Test results	Precision	0.9429	**0.9482**	0.9470	0.9269	0.8631
	Recall	0.8504	0.9433	**0.9481**	0.9352	0.5000
	F1-score	0.8804	0.9454	**0.9475**	0.9297	0.6108
Validation results	Precision	0.9427	**0.9480**	0.9466	0.9263	0.8627
	Recall	0.8499	0.9429	**0.9477**	0.9346	0.4997
	F1-score	0.8801	0.9451	**0.9471**	0.9292	0.6105

The best results are highlighted in bold

Experiments

Experiments were carried out using the following machine learning algorithms: Logarithmic Regression (LR), Random Forest (RF), K-nearest-neighbours (KNN) and AdaBoost with Decision Tree (AdaBoost). In addition, random class assignment was used as a baseline (Baseline). Data were split into 80% training data, 10% test data and 10% validation data. Since research publications are more prominent in the dataset than non-research publications with over 92.58%, the class distribution was stratified when splitting the data into training, test and validation set. In order to reduce the size of the dataset, publications from publishers with fewer than 5000 publications were removed. The reason for this decision was that dozens of publishers were found with only a minimal number of publications, in several cases with only one publication. Although the exclusion of these publishers does not have a substantial impact on the results achieved, it has helped to make the dataset clearer. Hyper-parameters have been optimised for each individual model by using grid search. The results of the experiments can be seen in Table 2. The performance of the individual models were calculated using the F1-score, as some models tend to have a high precision but a low recall value. The F1-score combines these precision and recall values into a new metric that allows for a harmonic mean between these two scores (for a full explanation see the Appendix). All metrics were calculated using the Python library *scikit-learn*.[11]

Results

Overall, the selected models perform very well on the dataset as can be seen in Table 2. The random forest classifier achieves the highest precision on both, the test dataset and the validation set, when comparing to the other algorithms with 94.8%. The k-nearest-neighbours algorithm achieves the highest recall and also the highest F1-score (both close to 95%), also on both, the test and validation set.

Table 3 shows the results of each model for the chosen classes research and non-research. As can be seen, the results are quite diverse as each model shows advantages and disadvantages. For example, Logarithmic Regression outperforms all other models when

[11] https://pypi.org/project/scikit-learn/

Table 3 Classifier results for each class based on test set

Algorithm						
	Measure	LR	RF	KNN	AdaBoost	Baseline
Research	Precision	**0.9924**	0.9761	0.9699	0.9536	0.9258
	Recall	0.8443	0.9618	0.9737	**0.9769**	0.4997
	F1-score	0.9124	0.9689	**0.9718**	0.9651	0.6491
Non-research	Precision	0.3209	0.5967	**0.6546**	0.5847	0.0740
	Recall	**0.9193**	0.7062	0.6228	0.4065	0.4998
	F1-score	0.4757	**0.6469**	0.6383	0.4796	0.1289

The best results are highlighted in bold

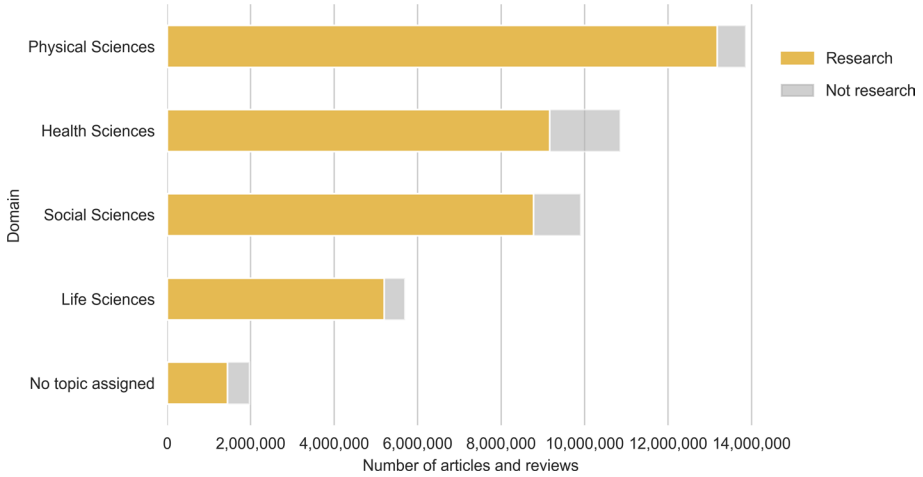

Fig. 1 Classifier results applied on domains in OpenAlex

classifying research contributions with over 99% precision. Further, it achieves the highest recall among the other models when predicting non-research publications. However, it performs poorly when detecting non-research publications (only about 32% precision). Also, in comparison with the other models, it achieves a lower recall with only about 84% when predicting the research class. The highest precision for predicting non-research publication is achieved by the k-nearest-neighbours algorithm (about 65%). Along with the highest F1-score for predicting the research class, the k-nearest neighbours algorithm was chosen for productive use.

Validation

In this section, the classifier will be applied and compared with real data. Firstly, the classifier will be compared with OpenAlex to see the impact of the classification results. Secondly, the classifier will be compared with Scopus. Finally, two small samples of one hundred publications each will be tested to evaluate the accuracy of the classifier.

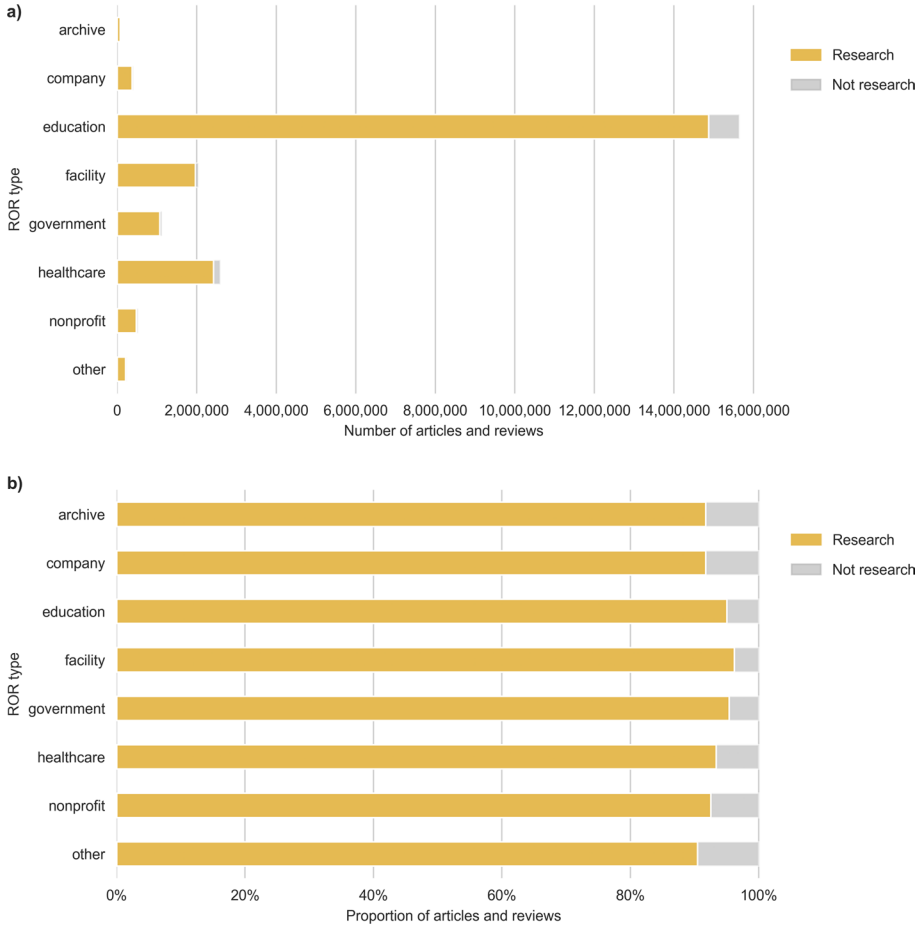

Fig. 2 Classifier results applied on institutional ROR types in OpenAlex. Only the related institution of the corresponding author was considered (information about the corresponding author is not always available in OpenAlex)

Comparison with OpenAlex

Overall, the classifier recognised 4,589,967 out of 42,701,863 *articles* and *reviews* from journals between 2014 and 2023 as non-research publications. This corresponds to about 10.75%. When applying the classification results to OpenAlex domains (Fig. 1), one can see that publications from the Health and Social Sciences were more frequently classified as non-research contributions than publications from Physical and Life Sciences. This is probably due to many case reports and abstracts in the Health Sciences which usually do not contain references and are often one or two pages long. This observation is also made when assigning the classification results to institutional types from the Research Organization Registry (ROR),[12] which are included in OpenAlex. Here, the institutional types

[12] https://ror.org

healthcare and education are more frequent than, for example, the institutional type government for non-research publications (see Fig. 2).

Comparison with Scopus

When evaluating the classifier on a shared corpus based on DOI matching between OpenAlex and Scopus and using solely the document type classification of Scopus, it shows that only 327,522 out of 22,338,918 *articles* and *reviews* in journals between 2014 and 2023 are labeled as non-research publications by the classifier. This corresponds to about 1.5%.

Figure 3 shows a comparison of the coverage of *articles* and *reviews* in journals restricted to the publication years 2014 to 2023 between OpenAlex, Scopus and the introduced classifier. While OpenAlex classifies more than 24,465,047 (95.87%) publications in the shared corpus as *articles* and *reviews*, Scopus only counts 22,691,562 (88.92%) items as research. The classifier is positioned between the two and counts 23,781,060 (93.19%) items as research.

Manual check

Two cases of samples were drawn and checked manually:

- Sample 1 (n=100): *Articles* and *reviews* in journals that are identified as research contribution by the classifier.
- Sample 2 (n=100): *Articles* and *reviews* in journals that are identified as non-research contribution by the classifier.

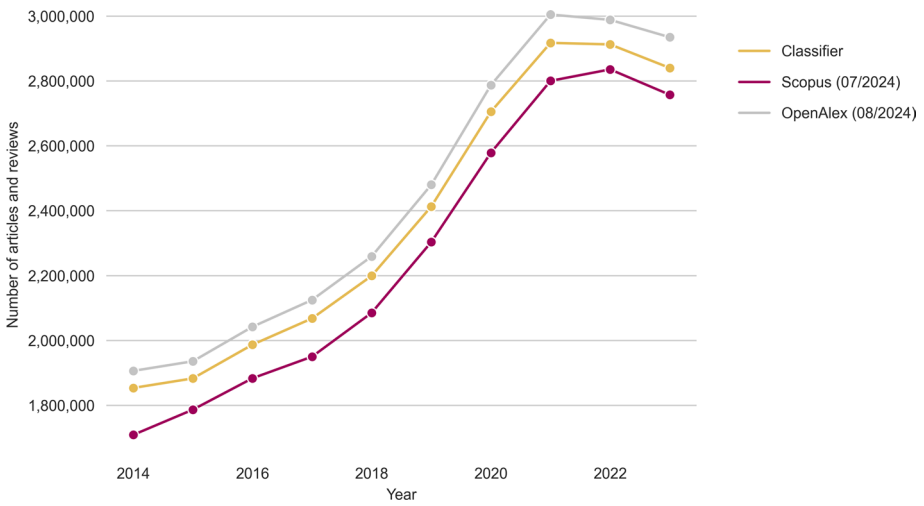

Fig. 3 Comparison of the classification of *articles* and *reviews* in journals between OpenAlex, Scopus and the introduced classifier based on a shared corpus

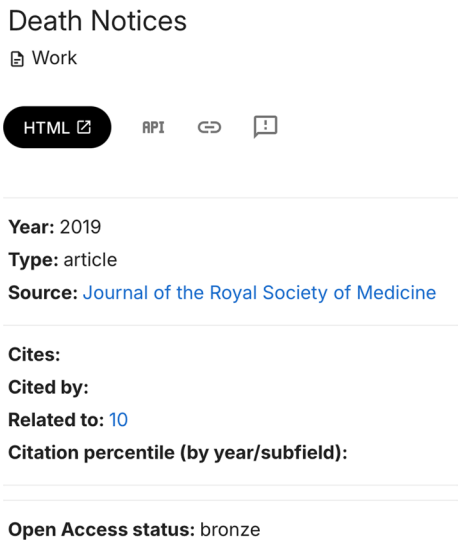

Fig. 4 Article that was correctly classified as non-research by the classifier (https://openalex.org/W4206081200)

Two equally sized samples were selected, even though the underlying populations are imbalanced (many more research contributions), because a typical application scenario for the classifier would be to select only one of two classes (e.g., selecting only journal items considered research by the classifier.)

Overall, 75% of items were correctly classified as research in Sample 1. 16% of items could not be checked due to not resolving DOIs or not loading websites. Also some items required a login. 9% of items were falsely classified as research. Some of them were *conference articles*,[13] *abstracts*, *corrections* or *book reviews*. For Sample 2, only *articles* and *reviews* that were classified by the classifier as non-research and were actually not an *article* or *review* (e.g. *editorial*, *abstract* or *paratext*) were considered as correct, whereas all *article* and *reviews* that were falsely classified by the model were considered as incorrect. Overall 69% of the results were classified correctly as non-research. In contrast, 21% were falsely classified as non-research. However, several of them originated from Malaysian, Indoniesian or Russian sources which could indicate a lack of provision of metadata by certain foreign (non-English) publishers or websites. Further 10% could not be checked, because resources were not found on the corresponding website or the corresponding website did not respond.

Examples of classified data

In this section, four examples of classified publications are provided: (a) an *article* that was correctly classified as non-research by the classifier (see Fig. 4), (b) an *article* that was erroneously classified as non-research by the classifier (see Figure 5), (c) an *article* that was correctly classified as research by the classifier (see Fig. 6) and (d) an *article*

[13] Conference articles were considered incorrect because technically they are not journal publications.

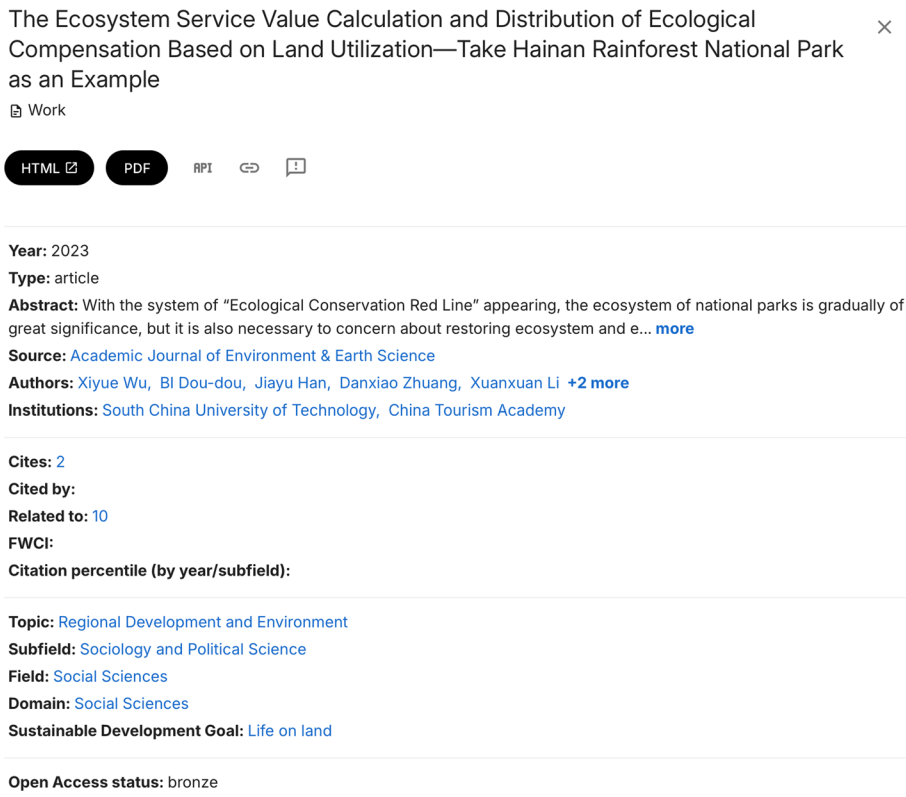

Fig. 5 *Article* that was erroneously classified as non-research by the classifier (https://openalex.org/works/w4380605217)

that was erroneously classified as research by the classifier (see Fig. 7). All examples were requested on October 6, 2025.

The examples shown here suggest that the classifier has particular difficulty correctly classifying publications with a low number of references. This is complicated by the fact that OpenAlex overlooks references in some cases (e.g. Fig. 5), whereby the publication has a reference count of 25 but OpenAlex only reports two references. However, there are also cases in which an editorial article contains at least one reference.[14]

Discussion and conclusion

The proposed classification approach provides a fast, scalable and simple method to assist the identification of research versus non-research publications in OpenAlex. With the help of the classifier, over 10% of *articles* and *reviews* can potentially be flagged as non-research contributions. In contrast, the classifier only detects around 1.5% of *articles* and *reviews* in Scopus as non-research based on a shared corpus between Scopus and OpenAlex.

[14] https://openalex.org/works/W4401071338

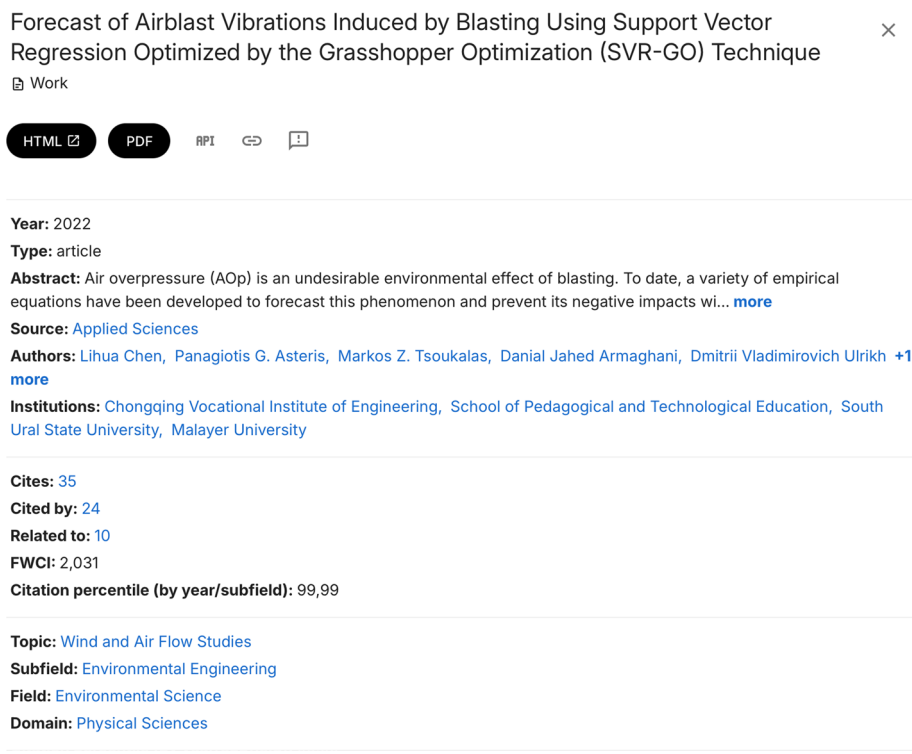

Fig. 6 *Article* that was correctly classified as research by the classifier (https://openalex.org/works/w4297988845)

The sample analysis shows that the classifier reaches around 89% accuracy for detecting research contributions and around 75% accuracy on classifying non-research works (when excluding resources that were not available, e.g. not resolving DOIs). This observation is in line with metrics that were achieved in the training process of the model. Compared to typical misclassification rates reported in databases such as Scopus and Web of Science, the error rates in this study are nonetheless quite large (Donner, 2017; Maisano et al., 2025, 2025).

The author would like to emphasise that the model performs well across works of different languages. Here, an even distribution was found in the classification of document types from English and non-English publications, although English publications are predominantly present in OpenAlex (Céspedes et al., 2025). In contrast, a language-based approach that relies on the title and abstract of a publication has the disadvantage that abstracts are often not available to the required extent and that a title cannot always be meaningful (if the title suggests a research article but is ultimately only an abstract). A full-text-based approach, in which content is extracted from PDFs, could be more helpful here (e.g. Caragea et al. (2016); Charalampous and Knoth (2017)). When comparing my approach with other existing methods, for instance the Leiden Ranking Open Edition, it shows that the classifier excludes fewer publications which can be attributed to the exclusion of journals and publications with missing metadata in the Leiden Ranking (Haupka, 2024; Van Eck

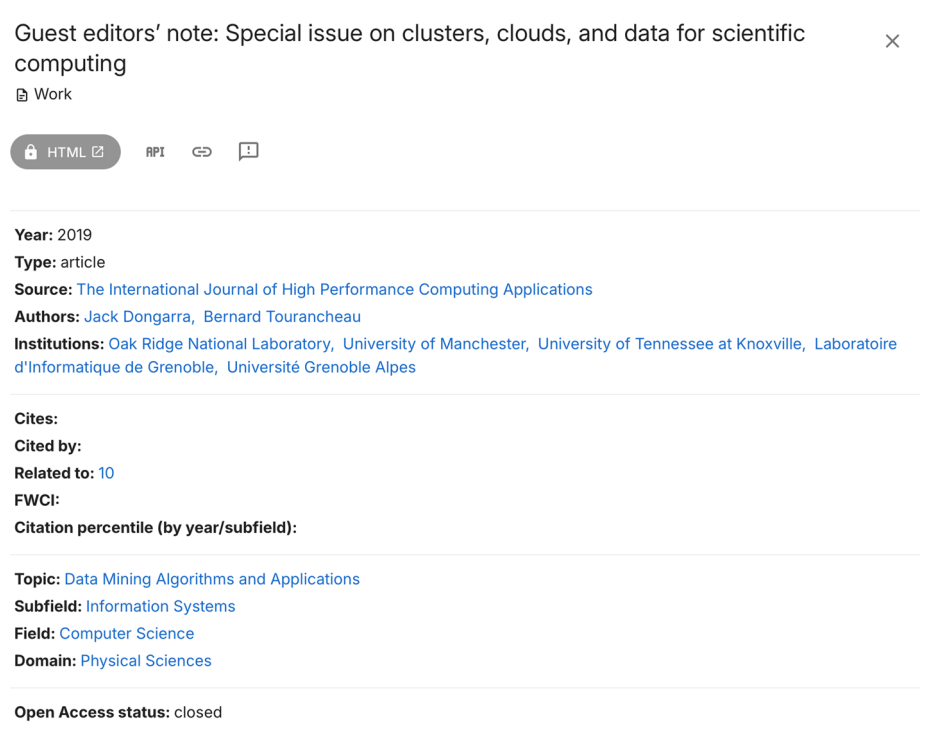

Fig. 7 *Article* that was erroneously classified as research by the classifier (https://openalex.org/works/w2568331913)

& Waltman, 2024)[15]. In contrast to semi-automated processes, where manual checks are necessary (e.g. Maisano et al., 2025), the classifier can also be applied to large amounts of data, making it more suitable for extensive analyses. However, the classifier can also be combined with manual validation. A noteworthy example of a platform for manual validation of metadata in OpenAlex is the Works-magnet developed by the French Ministry of Higher Education and Research.[16] It encourages users to correct automatically computed results in OpenAlex which then will be applied in the database. Mokhnacheva (2023) also suggests manual validation of document types using original information from publishers.

But my approach also comes with some limitations:

[15] The CWTS core classification is already implemented in OpenAlex (field name: *is_core*).

[16] https://works-magnet.esr.gouv.fr

- Coverage of metadata: significant discrepancies were noticed between metadata of different publishers and sources. While some publishers provide helpful metadata such as references and abstracts, others do not, which results in a lower accuracy of the classifier. Further, problems of missing metadata exist, e.g. in the cases of institutions (Zhang et al., 2024), page numbers (Delgado-Quirós & Ortega, 2024) and funding information (Culbert et al., 2025).
- Quality of metadata: for some publications, incorrect citation frequencies were observed. For example, several paratexts were found that have a high citation count although under investigation it shows that the citations are not correct.[17] Some page numbers are also not calculated correctly, either because OpenAlex contains incorrect page numbers or the page numbers are not available in a suitable format for calculation (e.g. containing letters).
- Quality of training set: although PubMed has a quality comparable to Scopus and Web of Science in the classification of document types, the author cannot guarantee that the training set does not contain errors. Additionally, the training set is based on a shared corpus between OpenAlex, Scopus, Web of Science, Semantic Scholar and PubMed, which can introduce a bias. For example, Web of Science is known to have a restrictive inclusion policy based, among other things, on document types. Thus, the classifier may work more accurately for publications that are indexed in most scholarly databases than, for example, for publications that are only included in OpenAlex.
- Classification approach: the proposed method only allows for binary classification (research vs. non-research). A clear differentiation between, for example, research articles and book reviews or research articles and editorial has not been considered.
- Selected features: the choice of features may be subject to bias. This also includes, for example, the treatment of missing page counts and the inclusion of citations in this study. However, the latter point in particular shows that the number of citations has no substantial impact on early publications as demonstrated in this blog post.[18] Yet it may be useful to include citation windows for the next iteration of the classifier.

Although the classifier was originally developed to recognise editorial content within a classified set of journal articles and reviews, it was noticed that many of the classification results indicate that OpenAlex also classifies a large number of *abstracts*, *case reports* and *book reviews* as journal articles. Here, the classifier can be used to identify journals that have a disproportionate share of non-research publications. This can help to check whether these journals have a focus on research articles or not. For example, the journal Reactions Weekly predominantly publishes *case reports* which however are labelled as articles in OpenAlex.[19] Another journal is the Bulletin of the Center for Children's Books, which mostly publishes *book reviews* that are classified as *articles* in OpenAlex.[20]

[17] https://openalex.org/works?page=1&filter=type%3Atypes%2Farticle,display_name.search%3Atitelei
[18] https://open-bibliometrics.de/posts/20250717-DocumentTypeClassifier/
[19] https://link.springer.com/journal/40278
[20] https://bccb.ischool.illinois.edu

Table 4 Mapping PubMed document types to the classes *research* and *non-research* Table is derived from Haupka et al. (2025)

	Types
Research	Cochrane Systematic Review, Systematic Review, Meta-Analysis, Review, Case Reports, Randomized Controlled Trial, Clinical Trial, Clinical Trial, Phase II, Clinical Trial, Phase III, Clinical Trial, Phase I, Clinical Trial, Phase IV, Controlled Clinical Trial, Pragmatic Clinical Trial, Journal Article, Comparative Study, Multicenter Study, Observational Study, Evaluation Study, Historical Article, Validation Study, Clinical Study, Randomized Controlled Trial, Veterinary, Twin Study, Clinical Trial, Veterinary, Classical Article, Observational Study, Veterinary, Corrected and Republished Article, Adaptive Clinical Trial, Evaluation Studies, Validation Studies, "Research Support, Non-U.S. Govt", "Research Support, N.I.H., Extramural", "Research Support, U.S. Govt, Non-P.H.S.", "Research Support, U.S. Govt, P.H.S.", Preprint, "Research Support, N.I.H., Intramural", "Research Support, American Recovery and Reinvestment Act", Equivalence Trial
Non-research	Published Erratum, Retraction of Publication, Retracted Publication, Editorial, News, Letter, Comment, Introductory Journal Article, Newspaper Article

Appendix

F1-score

The F1-score can be calculated as follows, where T_p is the number of true positives, F_p is the number of false positives and F_n is the number of false negatives.

$$F_1 = \frac{2 * T_p}{2 * T_p + F_p + F_n} \tag{1}$$

Mapping

Author contributions Conceptualization, investigation and writing—riginal draft: Nick Haupka.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was funded by the Federal Ministry of Education and Research (Grant Funding Number: 16WIK2301E, The OpenBib project).

Data availability The Jupyter notebook containing the source code analysis can be found on GitHub: https://github.com/naustica/openalex doctypes.

Declarations

Conflict of interest The author declares no conflict of interest.

References

Alperin, J. P., Portenoy, J., Demes, K., Larivière, V., & Haustein, S. (2024). *An analysis of the suitability of OpenAlex for bibliometric analyses*. Retrieved April 08, 2025, from arXiv:http://arxiv.org/abs/2404.17663

Bracco, L., Jeangirard, E., L'Hôte, A., & Romary, L. (2025). *How to build an Open Science Monitor based on publications? A French perspective*. Retrieved February 27, 2025, from arXiv:http://arxiv.org/abs/2501.02856

Caragea, C., Wu, J., Gollapalli, S., & Giles, C. (2016). Document type classification in online digital libraries. *Proceedings of the AAAI Conference on Artificial Intelligence, 30*(2), 3997–4002. https://doi.org/10.1609/aaai.v30i2.19075

Charalampous, A., & Knoth, P. (2017). *Classifying document types to enhance search and recommendations in digital libraries*. arxiv:1707.04134

Culbert, J. H., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2025). Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics, 130*(4), 2475–2492. https://doi.org/10.1007/s11192-025-05293-3

Céspedes, L., Kozlowski, D., Pradier, C., Sainte-Marie, M. H., Shokida, N. S., Benz, P., & Larivière, V. (2025). Evaluating the linguistic coverage of OpenAlex: An assessment of metadata accuracy and completeness. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24979

Delgado-Quirós, L., & Ortega, J. L. (2024). Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies, 5*(1), 31–49. https://doi.org/10.1162/qss_a_00286

Donner, P. (2017). Document type assignment accuracy in the journal citation index data of Web of Science. *Scientometrics, 113*(1), 219–236. https://doi.org/10.1007/s11192-017-2483-y

Haupka, N. (2024). *Identifying journal article types in OpenAlex*. https://subugoe.github.io/scholcomm_analytics/posts/oal_document_types_classifier/

Haupka, N., Culbert, J., Donner, P., Jahn, N., Lenke, C., Mayr, P. & Taubert, N. (2025). *Openbib: Selected curated open metadata based on OpenAlex.* Kompetenznetzwerk Bibliometrie. https://doi.org/10.5281/zenodo.15308680

Haupka, N., Culbert, J. H., Schniedermann, A., Jahn, N., & Mayr, P. (2025). *Analysis of the Publication and Document Types in OpenAlex, Web of Science, Scopus, Pubmed and Semantic Scholar*. arXiv:http://arxiv.org/abs/2406.15154

Haupka, N., Dörner, S., & Jahn, N. (2024). *Recent Changes in Document type classification in OpenAlex compared to Web of Science and Scopus*. https://subugoe.github.io/scholcomm_analytics/posts/openalex_document_types/

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies, 1*(1), 414–427. https://doi.org/10.1162/qss_a_00022

Jahn, N. (2025). Estimating transformative agreement impact on hybrid open access: A comparative large-scale study using Scopus. *Web of Science and Open Metadata: Scientometrics*. https://doi.org/10.1007/s11192-025-05390-3

Jahn, N. (2025). How open are hybrid journals included in transformative agreements? *Quantitative Science Studies, 6*, 242–262. https://doi.org/10.1162/qss_a_00348

Jeangirard, E. (2022). L'utilisation de l'apprentissage automatique dans le Baromètre de la science ouverte: une façon de réconcilier bibliométrie et science ouverte? *Arabesques, 107*, 10–11.

Jiao, C., Li, K., & Fang, Z. (2023). How are exclusively data journals indexed in major scholarly databases? An examination of four databases. *Scientific Data, 10*(1), 737. https://doi.org/10.1038/s41597-023-02625-x

Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A.. Weld, D.S. (2023). *The Semantic Scholar Open Data Platform*. arXiv:http://arxiv.org/abs/2301.10140

Maisano, D. A., Ferrara, L., Franceschini, F. (2025). Impact of Web of Science and Scopus Policies on Multiple Document-Type Classification. In *Proceedings of the 20th International Conference on Scientometrics & Informetrics (ISSI 2025)* (pp. 968–984). https://doi.org/10.51408/issi2025_053

Maisano, D. A., Mastrogiacomo, L., Ferrara, L., & Franceschini, F. (2025). A large-scale semi-automated approach for assessing document-type classification errors in bibliometric databases. *Scientometrics*. https://doi.org/10.1007/s11192-025-05244-y

Mokhnacheva, Y. V. (2023). Document types indexed in WoS and Scopus: Similarities, differences, and their significance in the analysis of publication activity. *Scientific and Technical Information Processing, 50*(1), 40–46. https://doi.org/10.3103/S0147688223010033

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., & Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ, 6*, e4375. https://doi.org/10.7717/peerj.4375

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. arXiv:http://arxiv.org/abs/2205.01833

Schmidt, M., Rimmert, C., Stephen, D., Lenke, C., Donner, P., & Gärtner, S.. Stahlschmidt, S. (2025). *The Data Infrastructure of the German Kompetenznetzwerk Bibliometrie: An Enabling Intermediary between Raw Data and Analysis*. *Quantitative Science Studies, 6(1), 1129-1146*. https://doi.org/10.1162/qss.a.20

Van Eck, N. J., & Waltman, L. (2024). *A methodology for identifying core sources and core publications in OpenAlex*(Tech. Rep.). Zenodo. Retrieved March 21, 2025, from https://zenodo.org/records/13879947

Zhang, L., Cao, Z., Shang, Y., Sivertsen, G., & Huang, Y. (2024). Missing institutions in OpenAlex: Possible reasons, implications, and solutions. *Scientometrics, 129*(10), 5869–5891. https://doi.org/10.1007/s11192-023-04923-y