Can editorial decisions impair journal recommendations? Analysing the impact of journal characteristics on recommendation systems

Anonymous Author(s)

ABSTRACT

Recommendation services for journals help scientists choose appropriate publication venues for their research results. They often use a semantic matching process to compare e.g. an abstract against already published articles. As these services can guide a researcher's decision, their fairness and neutrality are critical qualities. However, the impact of journal characteristics (such as the abstract length) on recommendations is understudied. In this paper, we investigate whether editorial journal characteristics can lead to biased rankings from recommendation services, i.e. if editorial choices can systematically lead to a better ranking of one's own journal. The performed experiments show that longer abstracts or a higher number of articles per journal can boost the rank of a journal in the recommendations. We apply these insights to an active, open-source journal recommendation system. The adaptation of the algorithm leads to an increased accuracy for smaller journals.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; *Adversarial retrieval*; • **Computing methodologies** → Machine learning approaches; *Ranking*; Natural language processing.

KEYWORDS

Adversarial attack, Data poisoning, Journal recommendation, Scientific publishing

ACM Reference Format:
Anonymous Author(s). 2024. Can editorial decisions impair journal recommendations? Analysing the impact of journal characteristics on recommendation systems. In *Proceedings of 18th ACM Conference on Recommender Systems '24 (RecSys '24).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

With the growing publication ecosystem, recommendation services for papers, journals or citations gain in importance. Journal recommendation systems, for instance, help scientists to choose a journal to publish their article based on e.g. title or abstract. However, with the growing influence of these services, the motivation for scientists and publishers to game these algorithms increases as well. Beel et al. [2] describe steps for *Academic Search Engine Optimisation* for scientists. There were cases where publishers submitted erroneous metadata to bibliometric databases resulting in increased citation counts [3]. This leads to the question of whether journal recommendation services are susceptible to manipulation of the ranking. While publishers or journal editors cannot directly influence which words a scientist uses, they could easily influence underlying editorial decisions like the title length. Submission guidelines already differ per journal. For instance, the maximum abstract length for the journal "Emerging Infectious Diseases" is 150 words, while the maximum for "Parasites and Vectors" is 350. At the same time, there are general trends towards longer titles [11] and abstracts [18].

There is little research on the weaknesses of journal recommendation services. Feng et al. [8] tested the impact of the journal size on their recommendation system; Gu et al. [9] investigated how recommendation algorithms can be manipulated from the user side. The general research area is described as data poisoning in which an attacker modifies the input (poisons the data) of a model during training or inference to impact certain metrics of the model (in this case the rank of a journal in the recommendations) [19].

In this paper, we investigate the following research questions:

- RQ1: Can journals' editorial decisions impact their ranking in journal recommendation services?
- RQ2: Are some recommendation algorithms more susceptible to these changes?
- RQ3: Do these results generalise and apply to different scientific fields?

In particular, we explore how title length, abstract length, and the number of articles in a journal influence four different journal recommendation algorithms. We conduct experiments with journals from two exemplary topic areas that resulted in a sufficiently large data set.

The paper is organised as follows. Related work is presented in Section 2, while the tested algorithms are explained in Section 3. The experiments, including the setup, results and mitigation, are described in Section 4. A conclusion is given in Section 5. The source code (without the full data set due to copyright) is published at https://anonymous.4open.science/r/Journal-recommendation-bias-26B4/.

2 RELATED WORK

There is a wide range of different algorithms for journal recommendation ranging from co-author networks and citations [14] to using the "Aims & Scope" section of journals [10]. In practice, most available services focus on semantic similarity based on title, abstract, and/or keywords [7]. They generally use content-based filtering. A literature review on publication venue recommendation is described by Ajmal & Muzammil [1].

RecSys '24, October 14–18, 2024, Bari, Italy

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

The research area of how machine learning methods can be influenced by manipulating their input data is called *data poisoning*. In their survey [19], Tian et al. create a taxonomy of data poisoning attacks by differentiating between untargeted, targeted, and backdoor attacks as well as attacks during the training or inference stage. During the training phase, attacks usually work by flipping the labels of certain samples (label flipping attack) or perturbing the samples themselves. For computer vision, this usually means changing images in ways that are not perceivable by humans. Not all attacks, however, rely on modifying the samples or the labels, e.g. batch-ordering poisoning attacks a model by reordering samples in batches which impacts model training [16].

There is little research on weaknesses to data poisoning in journal recommendation systems. Feng et al. [8] investigate the impact of different journal sizes on the accuracy of their recommendation system. They state that the accuracy for the first hit varies from 0.27 to 0.66 from smaller journals to bigger journals and rate these results as satisfactory. Gu et al. investigate weaknesses with respect to the input data [9]. They test if they can inject words into the user input to improve the rank of a specific journal. Inserting up to three words already shows a major impact on the system's accuracy.

Similarly to the classification of research papers, Descampe et al. [4] use data poisoning to misclassify the category of newspaper articles. They do so by changing or injecting a few words into the articles' text and successfully test the results for Word2Vec embeddings in combination with Naïve Bayes, Multi-Layer Perceptrons and a Recurrent Neural Network. Changing between 6% to 15% of the article usually suffices to misclassify the article to a target class.

3 SUPERVISED RECOMMENDATION ALGORITHMS

In contrast to related work, we test additional algorithms and manipulation scenarios that could be realised from the journal or editorial side. The following four supervised algorithms are tested. They are selected because they are currently used or have been used for recommendation services running in production. We also include the algorithm from the only prior work on adversarial attacks [9]. The algorithms use title and abstract or only the abstract. If they use both, they employ a weighted mechanism to combine the results for title and abstract (e.g. [6, 9]). To better compare the algorithms and reduce the amount of test cases, we train and test all algorithms with just the title in the long title scenario and just the abstract in the other two scenarios (see Section 4.1).

3.1 TF-IDF with Rocchio or kNN:

Gu et al. [9] use the Term Frequency - Inverse Document Frequency (TF-IDF) algorithm with the Rocchio classifier (also known as nearest centroid classifier) or the k-nearest neighbors algorithm (kNN) for journal recommendation and its manipulation. We use the same algorithm combination with the same cleaning and stemming steps for title and abstract; but we omit the article keywords, as some of the investigated algorithms do not use them and the information is unavailable via the Dimensions API.

3.2 Elasticsearch with sum or mean:

The recommendation services B!SON [6], JANE [15] and the one by Elsevier [13] use Lucene, Elasticsearch (which internally uses Lucene) or the same Okapi best-matching-25 algorithm (Okapi BM-25) for journal recommendation. Okapi BM-25 is similar to TF-IDF but introduces several constants for weighing the terms. All the aforementioned services look for similar articles based on the user's inputs and then sum or average the articles' scores per journal. B!SON considers the top 100 articles, Elsevier 1,000,000 and JANE 50. We test the algorithm for the top 50 and top 100 articles.

3.3 Journal embedding:

The Open Journal Matcher service [5] uses the concatenated abstracts to build an embedding per journal with the "en_core_web_md" model from the spaCy library. These embeddings are then compared with the embedding of the user input to find the best match. The comparison is performed by spaCy by calculating the cosine similarity of the averaged word vectors. We test this approach with both, the abstract and title of the articles.

3.4 fastText with CNN:

Both Son et al. [17] as well as the Pubmender recommendation service [8], use a combination of word embeddings together with a convolutional neural network (CNN) and fully-connected layers for journal recommendation. Son et al.'s network uses fastText embeddings [12], a one-dimensional convolutional layer with a window size of two and 1800 filters, max pooling with a kernel size of two, and two dense layers with the first one having 1000 dimensions. Additionally, they perform stop word removal. Feng et al. [8], employ three convolutional layers with max pooling, followed by three dense layers.

We used the same approach (including the stop word removal) as Son et al. [17] because they describe all parameters in detail. However, we reduced the network's size to a fourth (convolutional layer with 450 filters and dense layer with 250 dimensions) to achieve reliable convergence with our smaller data set.

4 EXPERIMENTS

The focus of our experiments is on the effect of sample injections for one label (journal) depending on different scenarios. We do not modify the article content itself.

4.1 Experimental setup

The attack scenarios are tested for two exemplary scientific fields which have numerous journals with a sufficiently large number of articles: mechanics of materials and infectious diseases. For each of these fields, 20 journals are selected from the Scimago Journal Ranking[1] based on their assigned subject categories and having at least around 1000 articles. The resulting journals are listed in the supplementary material.

The articles of these journals are fetched using the API of the bibliometric database Dimensions[2] by searching for the journal ISSN and validating that the title and abstract are present in the

[1] https://www.scimagojr.com/journalrank.php

[2] https://www.dimensions.ai/

Table 1: Results for the long abstract scenario. “SoR” is used as an abbreviation for “Shift of Rank”, “Acc” for “Accuracy”, and “SD” for “Standard deviation”.

Method	Engineering					Medicine				
	SoR	SD SoR	Acc Baseline	Acc Attack	SD Acc Attack	SoR	SD SoR	Acc Baseline	Acc Attack	SD Acc Attack
FastTextCNN	0.25	2.13	0.280	0.228	0.061	-1.779	2.701	0.222	0.156	0.019
Elasticsearch <sum,100>	-2.15	0.66	0.437	0.434	0.009	-4.312	1.192	0.360	0.328	0.017
Elasticsearch <average,100>	-0.62	0.31	0.218	0.243	0.005	-1.523	0.622	0.104	0.116	0.003
Elasticsearch <sum,50>	-1.79	0.53	0.433	0.442	0.007	-3.921	1.050	0.364	0.341	0.014
Elasticsearch <average,50>	-0.97	0.40	0.231	0.240	0.005	-2.220	0.728	0.118	0.123	0.004
Journal embedding	0.63	1.54	0.245	0.246	0.004	1.557	2.013	0.183	0.178	0.004
TF-IDF kNN	0.06	0.46	0.303	0.290	0.005	-1.049	0.895	0.274	0.269	0.004
TF-IDF Rocchio	0.34	1.17	0.362	0.364	0.004	-1.023	1.015	0.287	0.279	0.005

Table 2: Results for the long title scenario. “SoR” is used as an abbreviation for “Shift of Rank”, “Acc” for “Accuracy”, and “SD” for “Standard deviation”.

Method	Engineering					Medicine				
	SoR	SD SoR	Acc Baseline	Acc Attack	SD Acc Attack	SoR	SD SoR	Acc Baseline	Acc Attack	SD Acc Attack
FastTextCNN	0.65	2.84	0.272	0.241	0.015	1.288	2.284	0.242	0.200	0.009
Elasticsearch <sum,100>	-1.42	0.68	0.372	0.359	0.006	-1.934	0.856	0.260	0.242	0.007
Elasticsearch <average,100>	0.62	0.45	0.200	0.202	0.005	0.804	0.548	0.091	0.097	0.003
Elasticsearch <sum,50>	-0.67	0.56	0.379	0.366	0.004	-0.885	0.779	0.251	0.243	0.008
Elasticsearch <average,50>	0.17	0.52	0.181	0.197	0.004	0.422	0.593	0.090	0.090	0.004
Journal embedding	-0.56	0.82	0.249	0.252	0.002	0.054	1.776	0.200	0.189	0.005
TF-IDF kNN	0.64	1.03	0.164	0.164	0.007	0.134	1.231	0.151	0.149	0.004
TF-IDF Rocchio	-0.58	0.83	0.245	0.250	0.003	-0.919	0.992	0.194	0.198	0.005

Table 3: Results for the mega journal scenario. “SoR” is used as an abbreviation for “Shift of Rank”, “Acc” for “Accuracy”, and “SD” for “Standard deviation”.

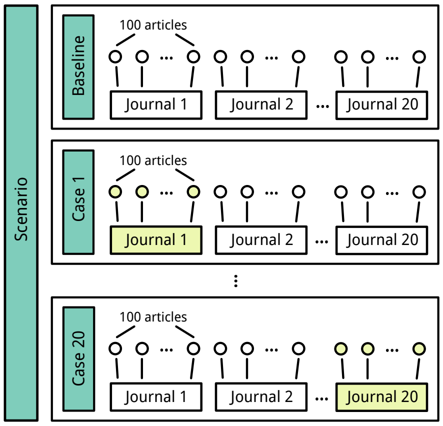
Method	Engineering					Medicine				
	SoR	SD SoR	Acc Baseline	Acc Attack	SD Acc Attack	SoR	SD SoR	Acc Baseline	Acc Attack	SD Acc Attack
FastTextCNN	-2.98	2.42	0.280	0.258	0.093	-2.851	2.962	0.188	0.210	0.056
Elasticsearch <sum,100>	-4.47	0.91	0.443	0.401	0.015	-4.741	0.442	0.390	0.350	0.016
Elasticsearch <average,100>	-1.27	0.52	0.233	0.215	0.006	-1.418	0.372	0.104	0.115	0.004
Elasticsearch <sum,50>	-3.91	1.04	0.440	0.421	0.011	-4.227	0.492	0.382	0.365	0.013
Elasticsearch <average,50>	-2.10	0.69	0.240	0.212	0.004	-2.326	0.416	0.122	0.110	0.004
Journal embedding	-0.32	0.68	0.261	0.238	0.002	-0.016	0.576	0.182	0.194	0.002
TF-IDF kNN	-4.74	1.02	0.316	0.296	0.016	-4.751	0.385	0.272	0.269	0.011
TF-IDF Rocchio	-1.12	0.54	0.386	0.383	0.003	-0.871	0.364	0.296	0.313	0.002

returned JSON. The test set consists of 100 articles per journal, thus 2000 articles in total per field and circa 10% of the article data.

To simulate different editorial decisions, we consecutively filter the articles of the training set for certain scenarios:

- **Mega journal**: One journal has significantly more articles than all other journals. The attacking journal has 300 articles, while the others have 100 articles.
- **Long titles**: One journal enforces longer titles meaning all of its articles have a specified minimum title length while the other journals have a maximum title length. Each journal has 100 articles. The attacking journal has titles longer than 150 characters, the others have titles with less than 100 characters.

Figure 1: A scenario with the baseline and 20 cases in each of which a different journal is modified (depicted in yellow).



- **Long abstracts**: The same as above but applied to the abstract instead of the title. In this case, the attacking journal has abstracts with more than 200 words while the others have abstracts with less than 200 words.

For each of these scenarios, we modify one journal to be the journal that performs the attack. Applying this to all 20 journals, we average the results (see Section 4.2). The boundaries of the thresholds are limited by the distribution in the data set. A wider distance for the abstract thresholds is not possible under the requirement of using at least 100 articles per journal for training. The overall experiment structure with the test scenarios and attack cases is presented in Figure 1.

The results are measured using the accuracy and *shift of rank* metrics. The accuracy tests whether the journal in which an article from the test set was originally published is correctly ranked first in the recommendation ("accuracy@1"). This is averaged for all articles in the test set and for the 20 different cases. Further, we introduce the *shift of rank (SoR)*, a metric that measures how the attacking journal ranks compared to the baseline case, in which no manipulation occurs. If the journal appears at rank 5 for a test input in the baseline case but at rank 2 for the same input in the manipulated case, the shift of rank would be −3 and the manipulation attempt would be successful. For both metrics, the standard deviation is calculated.

4.2 Results

The results in Tables 1, 2, and 3 provide evidence that the effect of data manipulation is generally stronger for the medicine field whereas we see overall higher accuracy in engineering. Moreover, embedding-based approaches seem more resistant to manipulation than the TF-IDF-based algorithms. Table 1 shows that Elasticsearch with summation is vulnerable for the long abstract scenario because it exhibits the strongest shift of ranks. However, the accuracy also drops significantly for the FastTextCNN algorithm. The algorithms are overall not severely impacted by the long title scenario in Table 2. The strongest effect on the shift of rank is again for Elasticsearch with summation. Curiously, the shift of rank for FastTextCNN is slightly positive here. The mega journal scenario in Table 3 impacts all algorithms except for journal embeddings. While the accuracy stays similar, the shift of rank is particularly strong for FastTextCNN, Elasticsearch with summation, and TF-IDF with kNN. The effect increases as the mega journal size increases. This is shown in the additional experiments in the supplementary material.

4.3 Mitigation strategies for using an example service

There are several starting points to limit the attacks. For the mega journal scenario, it is sufficient to limit the number of articles per journal or use the average for the aggregation in Elasticsearch and not summation. For the long abstract scenario, the length of the abstracts could be truncated. No additional steps are necessary for the long titles scenario as it did not impact the algorithm accuracy.

To analyse the actual impact of applying these mitigation suggestions, we test the effect of switching from summation to average in the B!SON recommendation service which is open source[3]. It uses the Directory of Open Acess Journal[4] as its primary data source which contains over 20,000 journals. We measure the effect with a data set of one article per journal. The measurements for title and abstract only use Elasticsearch, while the overall score uses both inputs and also takes the embedding-based neural network employed by B!SON into account. The recommendation accuracy for the lower quartile of the journals according to their size rises from 0.032 to 0.055 for the title, from 0.051 to 0.082 for the abstract and from 0.232 to 0.284 overall. The accuracy for the upper quartile drops from 0.165 to 0.097 for the title, 0.20 to 0.116 for the abstract and changes from 0.258 to 0.268 overall. The detailed results are shown in the Supplementary Materials.

5 CONCLUSIONS

In this paper, we investigated the impact of data manipulation on journal recommendation systems. The effects on commonly used algorithms were in some cases severe with shifts of up to -4.7 positions (RQ1). Specifically, Elasticsearch with summation is impacted by all scenarios. All algorithms except for journal embeddings were vulnerable to the mega journal scenario (RQ2). The results apply to the two distinct fields of medicine and engineering (RQ3).

With the increasing number of scientific articles and journals, the role of specialised recommendation systems will grow in importance. Our results indicate important challenges that need to be considered when designing such systems. Based on the open-source recommendation system B!SON, we tested one possible mitigation technique against data manipulation: using averaging for aggregation instead of summation. The adaptation improved the accuracy for smaller journals while lowering the accuracy for bigger journals. Other recommendation approaches discussed in research such as citation or author networks are not common in available production services and therefore not further analysed. The impact of e.g. limitations on the number of references on this type of system could be an interesting direction for future research, though it might be challenging to create adequate test scenarios.

[3] https://gitlab.com/TIBHannover/bison/
[4] https://doaj.org

REFERENCES

[1] Sahar Ajmal and Muteeb Bin Muzammil. 2019. PVRS: Publication Venue Recommendation System A Systematic Literature Review. In *International Conference on Computing Engineering and Design, ICCED 2019, Singapore, April 11-13, 2019*. IEEE, New York, NY, USA, 1–6. https://doi.org/10.1109/ICCED46541.2019.9161106

[2] Jöran Beel, Bela Gipp, and Erik Wilde. 2010. Academic Search Engine Optimization (aseo) Optimizing Scholarly Literature for Google Scholar & Co. *Journal of scholarly publishing* 41, 2 (2010), 176–190.

[3] Lonni Besançon, Guillaume Cabanac, Cyril Labbé, and Alexander Magazinov. 2023. Sneaked references: Cooked reference metadata inflate citation counts. https://doi.org/10.48550/arXiv.2310.02192

[4] Antonin Descampe, Clément Massart, Simon Poelman, François-Xavier Standaert, and Olivier Standaert. 2022. Automated news recommendation in front of adversarial examples and the technical limits of transparency in algorithmic accountability. *AI & SOCIETY* 37, 1 (March 2022), 67–80. https://doi.org/10.1007/s00146-021-01159-3

[5] Mark Eaton. 2022. open-journal-matcher - A journal recommender tool built on the Directory of Open Access Journals. https://github.com/MarkEEaton/open-journal-matcher (retrieved: 2024-05-07).

[6] Elias Entrup, Anita Eppelin, Ralph Ewerth, Josephine Hartwig, Marco Tullney, Michael Wohlgemuth, and Anett Hoppe. 2023. Comparing different search methods for the open access journal recommendation tool B!SON. *International Journal on Digital Libraries* (July 2023). https://doi.org/10.1007/s00799-023-00372-3

[7] Elias Entrup, Ralph Ewerth, and Anett Hoppe. 2023. A Comparison of Automated Journal Recommender Systems. In *Linking Theory and Practice of Digital Libraries*. Springer Nature Switzerland, Cham, Switzerland, 230–238. https://doi.org/10.1007/978-3-031-43849-3_20

[8] Xiaoyue Feng, Hao Zhang, Yijie Ren, Penghui Shang, Yi Zhu, Yanchun Liang, Renchu Guan, and Dong Xu. 2019. The Deep Learning–Based Recommender System "Pubmender" for Choosing a Biomedical Publication Venue: Development and Validation Study. *Journal of Medical Internet Research* 21, 5 (May 2019), e12957. https://doi.org/10.2196/12957

[9] Zhaoquan Gu, Yinyin Cai, Sheng Wang, Mohan Li, Jing Qiu, Shen Su, Xiaojiang Du, and Zhihong Tian. 2020. Adversarial Attacks on Content-Based Filtering Journal Recommender Systems. *Computers, Materials & Continua* 64, 3 (2020), 1755–1770. https://doi.org/10.32604/cmc.2020.010739

[10] Son T. Huynh, Nhi Dang, Dac H. Nguyen, Phong T. Huynh, and Binh T. Nguyen. 2023. FPSRS: a fusion approach for paper submission recommendation system. *Applied Intelligence* 53, 8 (April 2023), 8614–8630. https://doi.org/10.1007/s10489-022-04117-8

[11] Feng Kevin Jiang and Ken Hyland. 2023. Titles in research articles: Changes across time and discipline. *Learned Publishing* 36, 2 (2023), 239–248. https://doi.org/10.1002/leap.1498

[12] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017*, Vol. 2. Association for Computational Linguistics, Valencia, Spain, 427–431. https://aclanthology.org/E17-2068

[13] Ning Kang, Marius A. Doornenbal, and Robert J.A. Schijvenaars. 2015. Elsevier Journal Finder: Recommending Journals for your Paper. In *ACM Conference on Recommender Systems, RecSys 2015, Vienna Austria, September 16-20 2015 (RecSys '15)*. Association for Computing Machinery, New York, NY, USA, 261–264. https://doi.org/10.1145/2792838.2799663

[14] Tribikram Pradhan and Sukomal Pal. 2020. A hybrid personalized scholarly venue recommender system integrating social network analysis and contextual similarity. *Future Generation Computer Systems* 110 (Sept. 2020), 1139–1166. https://doi.org/10.1016/j.future.2019.11.017

[15] Martijn J. Schuemie and Jan A. Kors. 2008. Jane: suggesting journals, finding experts. *Bioinformatics (Oxford, England)* 24, 5 (March 2008), 727–728. https://doi.org/10.1093/bioinformatics/btn006

[16] I Shumailov, Zakhar Shumaylov, Dmitry Kazhdan, Yiren Zhao, Nicolas Papernot, Murat A Erdogdu, and Ross J Anderson. 2021. Manipulating SGD with Data Ordering Attacks. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., Red Hook, NY, USA, 18021–18032. https://proceedings.neurips.cc/paper/2021/hash/959ab9a0695c467e7caf75431a872e5c-Abstract.html

[17] Huynh Thanh Son, Huynh Tan Phong, and Nguyen Huu Dac. 2020. An efficient approach for paper submission recommendation. In *IEEE REGION 10 CONFERENCE, TENCON 2020, Osaka, Japan, November 16-19, 2020*. IEEE, New York, NY, USA, 726–731. https://doi.org/10.1109/TENCON50793.2020.9293909

[18] Mike Thelwall and Pardeep Sud. 2022. Scopus 1900–2020: Growth in articles, abstracts, countries, fields, and journals. *Quantitative Science Studies* 3, 1 (April 2022), 37–50. https://doi.org/10.1162/qss_a_00177

[19] Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. 2022. A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning. *Comput. Surveys* 55, 8 (Dec. 2022), 166:1–166:35. https://doi.org/10.1145/3551636

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009