Is the ranking of PubMed similar articles good enough? An evaluation of text similarity methods for three datasets

Mariana Neves[1] **Ines Schadock**[1] **Beryl Eusemann**[1]
Gilbert Schönfelder[1,2] **Bettina Bert**[1] **Daniel Butzke**[1]
[1]German Centre for the Protection of Laboratory Animals (Bf3R),
German Federal Institute for Risk Assessment (BfR), Berlin, Germany
[2]Institute of Clinical Pharmacology and Toxicology,
Charité - Universitätsmedizin Berlin, Berlin, Germany

Abstract

The use of seed articles in information retrieval provides many advantages, such as a longer context and more details about the topic being searched for. Given a seed article (i.e., a PMID), PubMed provides a pre-compiled list of similar articles to support the user in finding equivalent papers in the biomedical literature. We aimed at performing a quantitative evaluation of the PubMed Similar Articles based on three existing biomedical text similarity datasets, namely, RELISH, TREC-COVID, and SMAFIRA-c. Further, we carried out a survey and an evaluation of various text similarity methods on these three datasets. Our experiments considered the original title and abstract from PubMed as well as automatically detected sections and manually annotated relevant sentences. We provide an overview about which methods better perform for each dataset and compare them to the ranking in PubMed similar articles. While results varied considerably among the datasets, we were able to obtain a better performance than PubMed for all of them. Datasets and source codes are available at: https://github.com/mariananeves/reranking

1 Introduction

Tools for searching for relevant publications in the biomedical literature need to rank results with respect to their relevance to the user (Fiorini et al., 2018). However, different users are often interested in different aspects of the publications.

A study of the search logs from PubMed showed many interesting aspects of the user interaction with the tool (Islamaj Dogan et al., 2009). Since 80% of the viewed abstracts derived from the top 20, a good ranking algorithm is important. Further, queries are rather short (less than four tokens) and usually composed of a mix of semantic associations, including abbreviations, diseases, chemicals, author names, etc. However, in some situations, such as when searching for a particular research goal, it is a complex task to precisely define the search using just a couple of words.

PubMed similar articles allow a search based on seed articles, i.e., it provides a pre-compiled list of articles that are similar to the given seed article[1] (Lin and Wilbur, 2007). It is a valuable resource with many applications, e.g., for building clusters of articles (Boyack et al., 2020), entity networks (Lee et al., 2016), or similarity-based datasets (Brown et al., 2019; Butzke et al., 2020).

In comparison with keywords-based queries, seed articles provide a larger context, and potentially, more information about the subject being searched, e.g., details about the research goal, or long names for some abbreviations. Seed articles have been previously used for a variety of tasks in information retrieval, such as in the construction of bag of words (White, 2018), recommendation systems (Zhang et al., 2022), or in systematic reviews (Wang et al., 2022).

The PubMed similar articles can easily be queried with the Entrez Programming Utilities (E-utilities)[2]. The methods behind the Pubmed similar articles are based on a probabilistic topic-based model for content similarity (called "PMRA") (Lin and Wilbur, 2007). The developers of the function evaluated their method on the data of the TREC 2005 Genomics track (Hersh et al., 2005) and compared it to BM25 (Robertson et al., 1994).

In this manuscript we aimed at evaluating the ranking from PubMed similar articles for a set of available text similarity datasets. Our goal was to analyze the performance of this function in comparison with state-of-the-art algorithms for text similarity. We did neither train nor fine-tune any of the methods we used, i.e., we relied only on methods which either do not need to be trained (e.g., BM25)

[1] https://pubmed.ncbi.nlm.nih.gov/help/#computation-of-similar-articles

[2] https://www.ncbi.nlm.nih.gov/books/NBK25497/

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 133–144

or which were pre-trained ones (e.g., language models). Further, in order to be able to compare results to PubMed similar articles, we relied only on datasets based on PubMed. In our experiments, we considered the original abstracts in PubMed as well as automatically selected sections and manually annotated facets (relevant sentences).

We are only aware of three previous similar evaluations of the PMRA algorithm: (i) the RELISH database, in which the authors compared PMRA to BM25 and TF-IDF for a large collection of more than 180k articles and more than 3k seed articles (Brown et al., 2019); (ii) the SMAFIRA-c dataset, in which an evaluation was carried out for three seed articles (Butzke et al., 2020); and (iii) an evaluation for seven seed articles with the focus on the abstracts' sections (Neves et al., 2019). A couple of previous projects also carried out an evaluation on some of the datasets that we used (Medić and Šnajder, 2022; Mysore et al., 2022). However, they did not aim at evaluating the performance of the PMRA algorithm.

The contributions of our work are the following: (a) a short review of the various text similarity methods and datasets; (b) an evaluation of these methods on the three selected datasets, and based on a variety of semantic features; (c) a comparison of these methods to PubMed similar articles; and (d) making available the derived datasets, manual annotations, and source code of our experiments and evaluation.

2 Datasets

We selected datasets composed of PubMed abstracts and which included annotations with respect to relevance or similarity to a particular seed article. Therefore, we skipped datasets that did not comply with one or more of these conditions, such as CSFCube (Mysore et al., 2021), which is composed of publications from the areas of computational linguistics and machine learning. We considered each dataset in two settings: (a) as originally released by the authors, and (b) their overlap with the recent list of similar articles from PubMed. We describe the three selected datasets in detail below. An overview of the datasets is shown in Table 1 and in Figure 1.

RELISH. It is a large database in which more than 180k PubMed abstracts were validated in terms of similarity to a seed article (Brown et al., 2019). We utilized the dataset used in the development of the Aspire tool (Mysore et al., 2022), which is available for download[3]. It contains three levels of similarity and we mapped them as follows: similar (1 or 2), and not similar (0).

TREC-COVID. It is a dataset based on COVID-19 publications that was used in a series of evaluations for information retrieval (Roberts et al., 2020). We utilized the dataset used in the development of the Aspire tool (Mysore et al., 2022), which is available for download[4]. Since the articles were not associated with their corresponding identifiers in PubMed, we first matched them automatically, by querying their titles in PubMed, followed by a manual checking of many of them which were matched to either none or more than one PMID. After the mapping step, we noticed that many of the seed articles contained only a couple of articles in their list. Therefore, we kept only the seed articles with at least 50 articles in their list, thus obtaining a total of 33 seed articles. From these 33 seed articles, one had to be removed because it contained no similar articles among the candidates. The dataset contains three levels of similarity and we mapped them as follows: similar (1 or 2), and not similar (0). Our manually edited dataset is available in our GitHub repository.

SMAFIRA-c. It is a small dataset which contains four case studies from the area of alternative methods to animal experiments (Butzke et al., 2020). For each seed article, the authors retrieved the similar articles from PubMed and carried out an annotation based on the similarity of the research goal. We mapped the original labels as follow: (a) similar: equivalent "++", partially equivalent "+(+)" and noteworthy "+"; and (b) non similar: limbo "L" and not equivalent "-".

Evaluation sets based on PubMed similar articles (sa-eval)

We considered PubMed similar articles as a baseline for comparing the methods. For each of the datasets above, we retrieved the PubMed similar articles (around Nov/22 and Jan/23). We computed an overlap between articles contained in both lists, i.e., PubMed similar articles and the original dataset. The evaluation set derived of this overlap, hereafter called "sa-eval", was usually smaller

[3] https://figshare.com/articles/dataset/RELISH-Aspire/19425506

[4] https://figshare.com/articles/dataset/TRECCOVID-RF-Aspire/19425515

Datasets	No. seeds	eval		sa-eval	
		total	**similar**	**total**	**similar**
RELISH	1,618	60 [53;60]	45 [1;60]	28 [0;58]	21 [0;56]
TREC-COVID	32	58 [44;73]	34 [0;60]	58 [44;73]	34 [0;60]
SMAFIRA-c	4	99 [95:102]	12 [10;33]	72.5 [38;78]	10.5 [8;31]

Table 1: Overview of the datasets, the complete sets (eval) and the sets based on PubMed similar articles (sa-eval). The number of articles are shown in the following format: median [min;max].

than the original one, except for the TREC-COVID, which had all of its PMIDs included in the list of PubMed similar articles. The resulting sa-eval sets are shown in Figure 1, with more details in Table 1.

3 Methods

When selecting the methods for our experiments, we started with the ones considered in Mysore, Cohan, and Hope (2022), followed by adding some additional ones that we found while researching the literature. We did not consider models trained specifically on one of the datasets, which is the case of ASPIRE (Mysore et al., 2022). We describe our methods in this section and we split them into three parts: text processing (cf. 3.1), text representation (cf. 3.2), and ranking (cf. 3.3).

3.1 Text processing

Except when explicitly mentioned, for both the seed and the candidate article, we considered the title and abstract of the articles for all datasets, which we obtained from PubMed using the Entrez Programming Utilities (E-utilities)[5]. We concatenated the title and the abstract into a single text. For methods for which sentences were necessary, e.g., sentence embeddings, we split the text into sentences with Python SciSpacy[6]. Besides titles and abstracts, we also considered other kinds of text in our experiments, as discussed below.

Discourse elements. We considered selected sections of the abstracts, which we extracted using a tool trained on PubMed abstracts (Jin and Szolovits, 2018). We tagged only the abstract of the articles, and did not consider the titles. The tool returns the following discourse elements: background, objective, methods, results, and conclusions. For these experiments, we considered each discourse element separately, by concatenating the sentences that were tagged with each one of them, following their original order in the abstracts.

Facets. As carried out in Mysore, O'Gorman, McCallum, and Zamani (2021), we manually annotated the relevant sentences (i.e., facets) for the seed articles in the SMAFIRA-c dataset, given that it contains only four seed articles. Our goal was to evaluate whether the facets could improve the results. The annotation was carried out by three annotators (ann1, ann2, ann3) with a PhD either in veterinary medicine or biology. We did not require any previous experience with annotation. We automatically split the sentences of the abstract using Python SciSpacy and asked the annotators to select sentences which were part of the research goal. No further instruction was given with respect to the annotation process. We ran experiments with the facets from each annotator, but also considered a union and an intersection of all of them. For all experiments, we concatenated the manually selected sentences as a single text, following their original order in the abstracts.

3.2 Text representation

We describe the various methods that we considered in our experiments below.

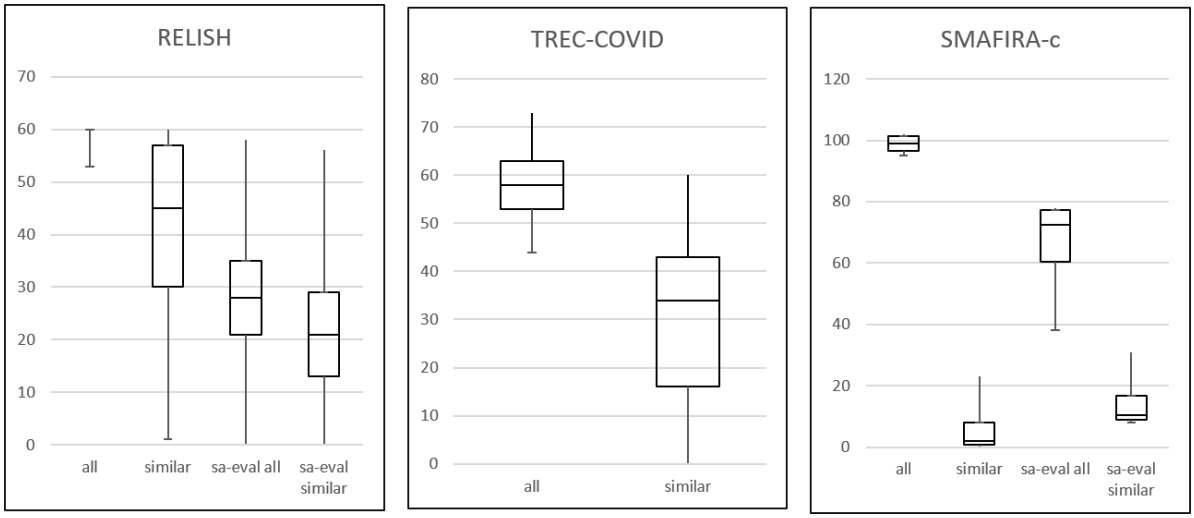
TF-IDF. We used the term frequency–inverse document frequency (TF-IDF) (Salton and Buckley, 1988) as implemented in Python Scikit-learn[7].

Word Embeddings. For the text similarity based on word embeddings, we vectorized the text using embeddings which are specific for the biomedical domain, namely: NCBI (Zhang et al., 2019), BioNLP-EVEX[8], Cambridge (Chiu et al., 2016), and ChemPatent (Zhai et al., 2019).

Pre-trained language models. For the sentence embeddings, given the automatically split sentences (cf. 3.1 above), we tokenized and vectorized them with each of the language models, which were

[5] https://www.ncbi.nlm.nih.gov/books/NBK25497/
[6] https://allenai.github.io/scispacy/
[7] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
[8] http://evexdb.org/pmresources/vec-space-models/

Figure 1: Variation of the number of abstracts per seed article and the corresponding number of similar articles. We show the complete datasets (on the left in each figure) and the sa-eval evaluation sets (on the right in each figure). Only one is shown for the TREC-COVID dataset since eval and sa-eval sets are the same.



the following: BioBERT (Lee et al., 2019), BLUEBERT (Peng et al., 2019), PubMedBERT (Gu et al., 2020), BioELECTRA (Kanakarajan et al., 2021), SPECTER (Cohan et al., 2020), SciBERT (Beltagy et al., 2019), and SimCSE (Gao et al., 2021) (both unsupervised and supervised models). We calculated the similarity between every pairwise combinations of sentences. The final score was given by the average of the scores of all pairwise combinations.

3.3 Ranking functions

For all text representations described above (cf. 3.2), we utilized the cosine similarity as implemented in Python Scikit-learn[9]. Next, we ranked the articles based on their similarity scores, from the most similar to the least similar. In addition to these, we tried one method which already includes its text representation, namely, Okapi BM25 algorithm (Robertson et al., 1994), as available in the rank-bm25 library[10].

Finally, for the sa-eval datasets, we evaluated re-ranking the candidates by considering two ranks: (i) the original one provided by PubMed similar articles (rank1), and (ii) the one returned by the corresponding method (rank2), as described above. We calculated the average between both ranks using three methods: (i) arithmetic average, i.e., $(rank1 + rank2)/2$, (ii) geometric average, i.e., $\sqrt{rank1 * rank2}$, and (iii) L2-Norm average, i.e., $\sqrt{rank1^2 + rank2^2}$.

4 Results

We show in this section the results that we obtained with our experiments. For the evaluation, we considered the metrics of Precision@20, Recall@20, R-Precision, and NDCG@20[11]. We considered the top 20 since this is the number of articles that most users usually screen during their search (Islamaj Dogan et al., 2009). We summarize the results in two tables, one for an overview of the highest scores (cf. Table 2) and one for an overview of the best performing methods (cf. Table 3).

Evaluation of the full datasets (eval sets). We first evaluated the full datasets (cf. first columns of Table 1). The best performance methods varied for the different metrics and datasets (cf. "E-Ta" row in Table 2). The recall was similar across the datasets, but precision, r-precision, and NDCG were much higher for the RELISH dataset. Indeed, the proportion of relevant articles is much higher for this dataset (cf. Figure 1).

Evaluation of the dataset based on similar articles (sa-eval). The results for the sa-eval datasets (S-Ta row in Table 2) were rather similar to the ones obtained for the full datasets, except for a few exceptions. The r-precision for the RELISH dataset was much lower than the ones for the "eval"

[9] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html

[10] https://pypi.org/project/rank-bm25/

[11] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.ndcg_score.html

Exp.	RELISH				TREC-COVID				SMAFIRA-c			
	r	p	r-p	ndcg	r	p	r-p	ndcg	r	p	r-p	ndcg
E-Ta	0.39	0.78	0.76	0.79	0.38	0.59	0.59	0.62	0.40	0.33	0.36	0.49
E-Sc	0.40	0.78	0.76	0.79	0.40	0.61	0.59	0.65	0.43	0.33	0.36	0.49
E-Fc	-	-	-	-	-	-	-	-	0.52	0.43	0.44	0.57
P-SA	0.37	0.73	0.48	-	0.39	0.59	0.56	-	0.52	0.39	0.43	-
S-Ta	0.37	0.74*	0.49*	-	0.38	0.59	0.59*	-	0.48	0.38	0.41	-
S-Sc	0.37	0.74*	0.49*	-	0.40*	0.61*	0.59*	-	0.51	0.40*	0.43	-
S-Rr	0.75*	0.51	0.58*	-	0.41*	0.63*	0.59*	-	0.40	0.33	0.33	-
S-Fc	-	-	-	-	-	-	-	-	0.54*	0.41*	0.48*	-

Table 2: Summary of the best scores for all experiments in terms of Recall@20 (r), Precision@20 (p), R-Precision (r-p), and NDCG@20 (ndcg). The first three rows correspond to experiments with the complete datasets (eval), when relying on title and abstracts (E-Ta), sections (E-Sc), and facets (E-Fc). The following row (P-SA) is the Pubmed similar articles (PMRA algorithm, baseline), as originally retrieved from PubMed. The last four rows correspond to the experiments with the modified datasets (sa-eval), when relying on title and abstracts (S-Ta), sections (S-Sc), re-ranking (S-Rr), and facets (S-Fc). Only the results below the P-SA row can be compared to the PubMed baseline. * indicates results that outperformed PubMed ranking (P-SA baseline).

Methods	RELISH		TREC-COVID		SMAFIRA-c		Total
	eval	sa-eval	eval	sa-eval	eval	sa-eval	
tfidf-cosine	R_{ta} P_{ta} RP_{ta} N_{ta} R_{sc} P_{sc} RP_{sc} N_{sc}	R_{ta} P^*_{ta} RP^*_{ta} R_{sc} P^*_{sc} RP^*_{sc}	P_{sc} N_{sc}	P^*_{sc}	RP_{ta} N_{ta} N_{sc} RP_{fc} N_{fc}	R_{ta} P_{ta} RP_{rr}	25
bm25					R_{sc}		1
w2v-ncbi						R_{rr}	1
w2v-bionlp						P^*_{fc}	1
w2v-cambridge							-
w2v-chempat							-
biobert					P_{ta} P_{sc} P_{fc}	R_{sc} P^*_{sc} RP_{sc}	6
pubmedbert					R_{ta}	RP_{ta}	2
bluebert			N^*_{ta} R_{sc}	R^*_{sc}		RP^*_{fc}	4
bioelectra							-
specter	R_{ta} P_{ta} P_{sc} RP_{sc}	R_{ta} P^*_{ta} R_{sc} P^*_{sc} R^*_{rr} P_{rr} RP^*_{rr}	P^*_{ta} RP^*_{ta} N^*_{ta} R_{sc} RP_{sc}	P_{ta} RP^*_{ta} R^*_{sc} RP^*_{sc} P^*_{rr} RP^*_{rr}	RP_{sc}		23
scibert		R_{sc}	P^*_{ta}	P_{ta}			3
simcse_sup		R_{sc}	R^*_{ta} R_{sc}	R_{ta} R^*_{sc} R^*_{rr} RP^*_{rr}	RP_{sc}	R^*_{fc}	9
simcse_usup		R_{sc}			P_{sc} R_{fc}	R_{rr} P_{rr} P^*_{fc} RP^*_{fc}	7

Table 3: Summary of the best scoring methods for all experiments. Upper cases letters refer to the metrics: r@20 (R), p@20 (P), r-prec (RP), and ndcg (N). Subscripts refer to the text processing or re-ranking: text from title and abstract (ta), sections (sc), re-ranking (rr), and facets (fc). * indicates results that outperformed PubMed ranking (PMRA algorithm).

dataset, and all results for the SMAFIRA-c dataset were sligtly higher. It should be noted that the annotation of SMAFIRA-c and the RELISH datasets were based on the PubMed similar articles available at the time. In comparison to the results from PubMed Similar articles (P-SA), only few scores were somewhat higher than these, namely, precision and r-precision for RELISH, and r-precision for TREC-COVID. We did not calculate the NDCG score for PubMed similar articles (P-SA), since the metric is based on similarity scores, e.g., probability or confidence values. While the PMRA algorithm returns scores for the similar articles, they are not suitable for such calculation.

Experiments with discourse elements (sections). Table 2 (E-Sc and S-Sc rows) shows only the best performing scores across all the sections. In general, we obtained a small improvement over the results based on the title and abstract, and more scores outperformed the ones from PubMed, also for the SMAFIRA-c dataset. Different sections obtained the best results for different datasets, and more than one were often equally good (cf. Table 4). "Background", "objective", and "conclusion" were the ones which obtained the best results, while "methods" and "discussion" were rarely the best performing ones.

eval	r@20	p@20	r-p	ndcg
RELISH	BO	BO	BO	BO
TREC-COVID	BC	C	O	C
SMAFIRA-c	B	BOC	BO	O
sa-eval	r@20	p@20	r-p	-
RELISH	all	all	BO	-
TREC-COVID	BC	C	O	-
SMAFIRA-c	B	B	C	-

Table 4: Summary of the sections that best performed for each dataset, also for eval and sa-eval. The sections are represented by their first letter: background (B), objective (O), methods (M), results (R), and conclusions (C). "all" means all of the five sections, i.e., BOMRC.

Re-ranking. Table 2 (S-Rr row) presents the results based on both PubMed and the methods' ranks and based on the three averages (cf. Section 3.3). The results were much worse for SMAFIRA-c, but, in general, a minor improvement was achieved for the other datasets. Two exceptions were the recall and precision for the RELISH dataset: the first had a large improvement, while the latter decreased. In general, the geometric average was the best performing one, followed by the arithmetic average in some cases (results not shown).

eval	r@20	p@20	r-p	ndcg
title-abs	0.40	0.33	0.36	0.49
ann1	0.44	0.36	0.36	0.57
ann2	0.44	0.43	0.44	0.55
ann3	0.46	0.36	0.42	0.54
union	0.46	0.36	0.41	0.56
intersection	0.52	0.40	0.41	0.51
sa-eval	r@20	p@20	r-p	-
pubmed	0.52	0.39	0.43	-
title-abs	0.48	0.38	0.41	-
ann1	0.54*	0.40*	0.45*	-
ann2	0.53*	0.41*	0.48*	-
ann3	0.51	0.39	0.48*	-
union	0.54*	0.40*	0.45*	-
intersection	0.53*	0.41*	0.44*	-

Table 5: Results using the original text based on the title and abstracts (title-abs), and based on the selected text annotated by each annotator (ann1, ann2, ann3), as well as their union and intersection. * indicates results that outperformed PubMed ranking (PMRA algorithm).

Experiments with facets. We asked three experts to annotate the four seed articles of SMAFIRA-c with respect to the research goal (cf. Section 3.1). For a total of 51 sentences (sum from all four reference articles), the kappa coefficient and level of agreement (McHugh, 2012) were the following: 0.51 (weak) between ann1-ann2, 0.25 (minimal) between ann1-ann3, and 0.33 (minimal) between ann2-ann3. Table 2 (E-Fc and S-Fc rows) shows that it brought an improvement for SMAFIRA-c, for both the complete (eval) and modified (sa-eval) datasets. Further, the scores were higher than the ones from the PMRA algorithm.

We experimented with facets from each of the annotators, as well as a union and intersection of their annotations. For the union, we considered any sentence that one of the annotator had selected, while the intersection consisted of the sentences that were annotated by at least two of them. Table 5 shows the results for each of these approaches, for both evaluation sets (eval and sa-eval). None of the results was lower than the ones obtained with the title and abstract, and many of them outperformed the PMRA algorithm.

5 Discussion

In this section we discuss some interesting points of our results, such as the best performing methods and a comparison between the datasets, as well as an error analysis.

Best performing methods. Table 3 presents the overview of the best performing methods for each dataset, for both evaluation sets (eval and sa-eval), and for the various text representations that we considered. On the one hand, some few methods never obtained the best performance, e.g., some of the word2vec embeddings and BIOELECTRA language model. On the other hand, two methods were the best performing ones, curiously, a simple one (TF-IDF), and a more complex recent one (SPECTER). While TF-IDF performed better for the RELISH and SMAFIRA-c datasets, SPECTER obtained the best results for the RELISH and TREC-COVID datasets. We could not observe that the best results for a particular type of experiment (e.g. facets or section) or metric (e.g., precision or recall) were correlated with some particular methods.

Different methods performed better for the different datasets. The best results for the RELISH dataset was restricted to two methods, namely TF-IDF and SPECTER. The ones based on discourse elements performed better with TF-IDF, while the ones with re-ranking averages were based on SPECTER. For the TREC-COVID datsets, best results involved many methods, most of them on the bottom part of Table 3, i.e., methods based on pre-trained language models. Finally, the best results for SMAFIRA-c were spread all over the table, and included the only best results obtained by BM25 and word2vec embeddings.

Comparison between the datasets. The three datasets are very different to each other, thus results cannot be compared between them. However, together, they provide a good overview of the performance of the available methods. On the one hand, the results that we obtained might be more reliable for datasets with more seed articles (i.e., RELISH), than with shorter ones (i.e. SMAFIRA-c). On the other hand, the RELISH dataset has a higher rate of similar articles (cf. Table 1 and Figure 1) than SMAFIRA-c. The TREC-COVID dataset lies between the two of them, having a higher number of seed articles than SMAFIRA-c, and a lower rate of similar articles than RELISH.

Two of the datasets (RELISH and SMAFIRA-c) relied on the output of the PubMed similar articles during their annotation. Indeed, we observed higher scores for SMAFIRA-c for the sa-eval evaluation set. However, the same did not occur for neither the RELISH nor the TREC-COVID datasets.

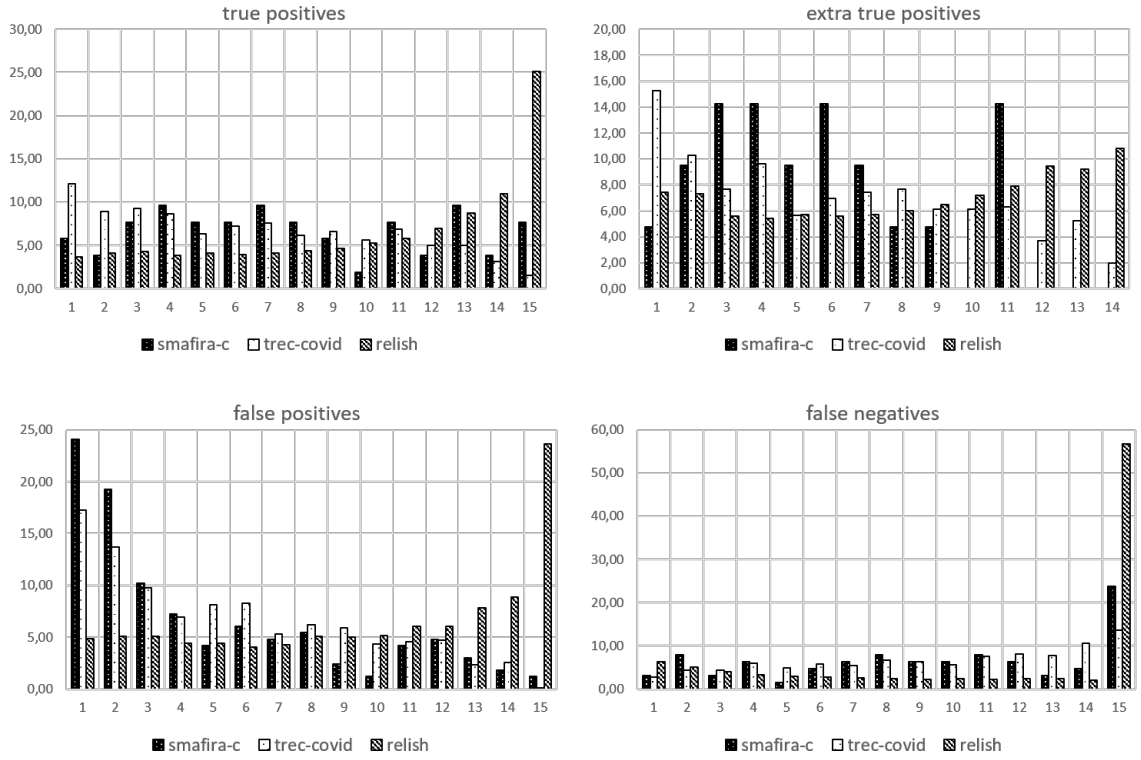
In order to reduce the high rate of similar articles in some of the datasets, we ran additional experiments by removing seed articles with such high rates from the evaluation set. We considered three values of rates (0.5, 0.25, and 0.1) and automatically removed seed articles with a rate of similar articles higher than these values. The resulting number of seed articles for the three rates were the following, respectively: 406, 114, 30 for RELISH, 13, 7, 5 for TREC-COVID, and 4, 3, 0 for SMAFIRA-c. We show our results in Appendix A.1. As expected, we noticed a decrease of the precision, since it is harder to find a similar article among the many candidates, and an increase in the recall, since there are less similar articles to be ranked.

Annotation of the facets. The annotation of the relevant sentences varied among the annotators (cf. Appendix A.2). While two of the annotators (ann1 and ann2) usually obtained similar annotations, one of them (ann3) annotated less sentences. However, annotations from any of them obtained good results, as presented in Table 5, even though it is a very subjective task, and though we practically provided no guidelines to the annotators. Further, it is not a very time-consuming task, since only the seed articles need to be annotated, thus allowing the integration of such type of input into search tools.

In general, sentences with more agreement between the annotators described the hypothesis or results of the experiments, usually with expressions such as “the aim of the study” or “we hypothesized”. The sentences that were selected by only one annotator comprised some additional details to the experiments, e.g., mutations in particular genes. In many cases, the sentences also contained details about the experimental models, i.e., an animal experiment, which was not relevant to the definition of similarity in the SMAFIRA-c dataset, nor did it belong to the research goal. However, this did not seem to compromise the results.

Error analysis. We ran an error analysis to detect interesting aspects of the predictions that we obtained from the various methods. For this analysis,

Figure 2: Visualizations of the number of methods (x-axis), including Pubmed PMRA, and the number of true positives, extra true positives, false positives, and false negatives (y-axis) that they returned. The graphic for extra TPs does not consider Pubmed PMRA, thus only 14 methods in the x-axis.



we considered the top 20 predictions of all methods (including the PMRA algorithm) and compiled the usual true positives (TPs), false positives (FPs), and false negatives (FNs). In addition to these, we also checked the TPs which were only predicted by the other methods, but not by the PMRA algorithm (hereafter called "extra TPs"). We plotted these analyses in Figure 2.

While any of the methods could predict many TPs for the RELISH dataset (cf. high gray bar for value "15" on the right), this does not occur for the TREC-COVID dataset, for which relatively high bars can be seen by low values (cf. white bars for values "1" to "4"). This means that, sometimes, just a couple of the methods were able to place relevant PMIDs in the top 20. Further, extra TPs, i.e., TPs in addition to the ones placed on the top 20 by the PMRA algorithm, are usually only found by some few methods (cf. high white and black bars for values up to "6"), thus proving that some methods are indeed better than others for particular datasets, and particularly for SMAFIRA-c and TREC-COVID.

On the one hand, the graphic for FPs shows that, for the SMAFIRA-c and TREC-COVID datasets, the highest rates of these errors are due by a couple of methods (cf. black bars on the left of the graph). On the other hand, many of the FPs for the RELISH dataset is common by all of the methods (cf. high grey bar on the right of the graph). Regarding the FNs, many of the mistakes are made by all of the methods (cf. the three highest bars on the right of the graph).

6 Conclusions

We presented an overview of available methods for the task of text similarity and provided a comprehensive evaluation for three datasets from the biomedical domain. We considered only datasets derived from PubMed and, in our experiments, evaluated various kinds of inputs, such as the title and the abstract of the articles as well as some (potentially relevant) parts of the abstracts, which have been either manually or automatically selected. We provided a variety of evaluations, based on the complete datasets, as well as a modified version of them for a comparison to the ranking algorithm from PubMed similar articles.

The results were diverse and the best performing methods were different for each of the datasets,

and for the various experiments that we carried out. However, both former and recent methods, e.g., TF-IDF and SPECTER, respectively, obtained good results for some of the datasets. Further, for all datasets, at least one of our experiments could outperform the ranking algorithm from PubMed similar articles. We hope that our survey can support the researchers when deciding about the most appropriate method to be used for a particular situation, based on either the dataset or hardware availability.

Limitations

Our experiments only used the abstracts of the articles and we did not consider the full text. Actually, the full text might provide valuable information for the decision about the similarity, while making it harder at the same time, since the text is much longer. However, full text is usually not available for many of the articles in PubMed, which would result in even shorter datasets.

We did not train any model based on the available datasets, as carried out by (Mysore et al., 2022). Indeed, our goal was to evaluate the dataset based on general-purpose text similarity algorithms that had not been fine-tuned for a particular dataset.

For the annotation of the facets, we simply asked the annotators to select sentences that were part of the research goal. However, it might not have been carried exactly as in the annotation of the SMAFIRA-c dataset.

Ethics Statement

Our experiments relied on published datasets derived from abstracts from PubMed. The various definitions of similarity in these datasets are solely based on scientific aspects of the abstracts and should not raise any ethical concerns.

References

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

Kevin W. Boyack, Caleb Smith, and Richard Klavans. 2020. A detailed open access model of the pubmed literature. *Scientific Data*, 7(1):408.

Peter Brown, RELISH Consortium, and Yaoqi Zhou. 2019. Large expert-curated database for benchmarking document similarity detection in biomedical literature search. *Database*, 2019. Baz085.

Daniel Butzke, Nadine Dulisch, Sebastian Dunst, Matthias Steinfath, Mariana Neves, Brigitte Mathiak, and Barbara Grune. 2020. Smafira-c: A benchmark text corpus for evaluation of approaches to relevance ranking and knowledge discovery in the biomedical domain.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174, Berlin, Germany. Association for Computational Linguistics.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Nicolas Fiorini, Robert Leaman, David J. Lipman, and Zhiyong Lu. 2018. How user intelligence is improving pubmed. *Nature Biotechnology*, 36(10):937–945.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

William Hersh, Aaron Cohen, Jianji Yang, Ravi Teja Bhupatiraju, Phoebe Roberts, and Marti Hearst. 2005. Trec 2005 genomics track overview. *NIST Special Publication*. 14th Text REtrieval Conference, TREC 2005 ; Conference date: 15-11-2005 Through 18-11-2005.

Rezarta Islamaj Dogan, G. Craig Murray, Aurélie Névéol, and Zhiyong Lu. 2009. Understanding PubMed® user search behavior through log analysis. *Database*, 2009. Bap018.

Di Jin and Peter Szolovits. 2018. Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3100–3109, Brussels, Belgium. Association for Computational Linguistics.

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. BioELECTRA:pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Kyubum Lee, Wonho Shin, Byounggun Kim, Sunwon Lee, Yonghwa Choi, Sunkyu Kim, Minji Jeon, Aik Choon Tan, and Jaewoo Kang. 2016. HiPub: translating PubMed and PMC texts to networks for knowledge discovery. *Bioinformatics*, 32(18):2886–2888.

Jimmy Lin and W. John Wilbur. 2007. Pubmed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics*, 8(1):423.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282. 23092060[pmid].

Zoran Medić and Jan Šnajder. 2022. Large-scale evaluation of transformer-based article encoders on the task of citation recommendation.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-vector models with textual guidance for fine-grained scientific document similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4453–4470, Seattle, United States. Association for Computational Linguistics.

Sheshera Mysore, Tim O'Gorman, Andrew McCallum, and Hamed Zamani. 2021. CSFCube - a test collection of computer science research articles for faceted query by example. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Mariana Neves, Daniel Butzke, and Barbara Grune. 2019. Evaluation of scientific elements for text similarity in biomedical publications. In *Proceedings of the 6th Workshop on Argument Mining*, pages 124–135, Florence, Italy. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets.

Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. 2020. TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19. *Journal of the American Medical Informatics Association*, 27(9):1431–1436.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *TREC*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513–523.

Shuai Wang, Harrisen Scells, Justin Clark, Bevan Koopman, and Guido Zuccon. 2022. From little things big things grow: A collection with seed studies for medical systematic review literature search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3176–3186, New York, NY, USA. Association for Computing Machinery.

Howard D. White. 2018. Bag of works retrieval: Tf*idf weighting of works co-cited with a seed. *International Journal on Digital Libraries*, 19(2):139–149.

Zenan Zhai, Dat Quoc Nguyen, Saber Akhondi, Camilo Thorne, Christian Druckenbrodt, Trevor Cohn, Michelle Gregory, and Karin Verspoor. 2019. Improving chemical named entity recognition in patents with contextualized word embeddings. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 328–338, Florence, Italy. Association for Computational Linguistics.

Li Zhang, Wei Lu, Haihua Chen, Yong Huang, and Qikai Cheng. 2022. A comparative evaluation of biomedical similar article recommendation. *Journal of Biomedical Informatics*, 131:104106.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific Data*, 6(1):52.

A Appendix

A.1 Results for lower rates of similar articles

Table 6 shows results with the various values of rate of similar articles that we considered.

A.2 Annotation of the facets

We show the annotations from each annotator, as well as the union and intersection of these, for the seed articles 16850029 (Table 7), 19735549 (Table 8), 21494637 (Table 9), and 24204323 (Table 10).

RELISH	r@20	p@20	r-p	ndcg
all	0.39	0.78	0.76	0.79
0.5	0.48	0.47	0.47	0.57
0.25	0.54	0.24	0.32	0.49
0.1	0.61	0.12	0.26	0.46
TREC-COVID	r@20	p@20	r-p	ndcg
all	0.38	0.59	0.59	0.62
0.5	0.37	0.24	0.27	0.31
0.25	0.39	0.09	0.13	0.24
0.1	0.32	0.05	0.10	0.19
SMAFIRA-c	r@20	p@20	r-p	ndcg
all	0.40	0.33	0.36	0.49
0.5	0.40	0.33	0.36	0.49
0.25	0.41	0.23	0.29	0.45
0.1	-	-	-	-

Table 6: Best results for the complete dataset (all) and modified version of the datasets with lower rates of similar articles.

16850029	ann1	ann2	ann3	U	I
1	x			x	
2					
3					
4	x	x	x	x	x
5					
6					
7		x		x	
8					
9					
10					
11	x			x	
12					
13					
14					
15	x	x		x	x

Table 7: Facets annotation for PMID 16850029, for each of the sentences in its abstract. We show the annotations by the three annotators, as well as union (U) and intersection (I) of these.

19735549	ann1	ann2	ann3	U	I
1	x	x		x	x
2					
3	x			x	
4	x	x	x	x	x
5					
6	x	x		x	x
7					
8					
9					
10	x			x	
11	x			x	
12		x		x	
13					
14	x	x		x	x
15	x	x	x	x	x

Table 8: Facets annotation for PMID 19735549, for each of the sentences in its abstract. We show the annotations by the three annotators, as well as union (U) and intersection (I) of these.

21494637	ann1	ann2	ann3	U	I
1	x	x		x	x
2	x			x	
3					
4					
5	x	x	x	x	x
6	x	x		x	x
7					
8	x	x	x	x	x

Table 9: Facets annotation for PMID 21494637, for each of the sentences in its abstract. We show the annotations by the three annotators, as well as union (U) and intersection (I) of these.

24204323	ann1	ann2	ann3	U	I
1	x	x		x	x
2	x			x	
3	x	x	x	x	x
4					
5					
6		x		x	
7		x		x	
8					
9	x			x	
10					
11					
12					
13	x	x		x	x

Table 10: Facets annotation for PMID 24204323, for each of the sentences in its abstract. We show the annotations by the three annotators, as well as union (U) and intersection (I) of these.