Appendix overview

This section provides an overview of the supplementary materials for our submission.

Appendix A offers an extended literature review encompassing citation screening datasets, evaluation measures used, and dataset coverage for other systematic literature review (SLR) steps. Appendix B presents detailed descriptions of the visualisations we have created. Appendix C provides documentation for the CSMED meta-dataset, including the datasheet. For the CSMED-FT dataset, please refer to Appendix D for its detailed documentation. In Appendix E, we delve into the specifics of dataset overlap. Appendices F and G provide comprehensive experimental details and expanded results for the baseline experiments conducted on the CSMED and CSMED-FT datasets.

All data loaders and data preprocessing scripts for CSMED are available under the following URL: `https://github.com/WojciechKusa/systematic-review-datasets`. CSMED-FT can also be accessed under the following URL: `https://github.com/WojciechKusa/systematic-review-datasets/raw/main/data/CSMeD/CSMeD-FT.zip` Additionally, the code to reproduce our experiments can be found at: `https://github.com/WojciechKusa/CSMeD-baselines`.

A Detailed literature review of datasets

We base our literature review on three recent surveys, which we extend to cover the results until May 2023:

- Systematic review conducted by O'Mara-Eves et al. [60] in 2015.
- Update to the review above, completed by Norman [56] in 2020.
- Systematic review conducted by van Dinter et al. [78] in 2021.

A.1 Citation screening datasets

We searched Google Scholar and Semantic Scholar for publications introducing new datasets for the citation screening task. We then searched for the forward citations of the original publication to find usages of the datasets. From our list, we excluded private datasets used in only one publication. We found 12 datasets fulfilling the criteria.[9] Table 6 presents a summary of these datasets.

Table 6: Systematic literature review datasets with their characteristics, sorted by the publication year. We included all publicly available datasets and private datasets which were used in more than one publication.

	Publication	# reviews	Domain	Data URL	Publicly available	In CSMeD
(1)	Cohen et al. [11], 2006	15	Drug	URL	✓	✓
(2)	Wallace et al. [82], 2010	3	Clinical	URL	✓	✓
(3)	Miwa et al. [54], 2014	4	Social science	—	—	—
(4)	Howard et al. [27], 2016	5	Mixed	URL	✓	✓
(5)	Scells et al. [71], 2017	93	Clinical	URL	✓	✓
(6)	Kanoulas et al. [30], 2017	50	DTA	URL	✓	✓
(7)	Kanoulas et al. [31], 2018	30	DTA	URL	✓	✓
(8)	Kanoulas et al. [32], 2019	49	Mixed	URL	✓	✓
(9)	Alharbi and Stevenson [2], 2019	25	Clinical	URL	✓	✓
(10)	Parmar [63], 2021	6	Biomedical	—	—	—
(11)	Wang et al. [88], 2022	40	Clinical	URL	✓	—
(12)	Hannousse and Yahiouche [22], 2022	7	Comp. Science	URL	✓	✓

A dataset created by Cohen et al. [11] containing 15 SLRs is the first and, up until today, one of the most commonly used to evaluate the effectiveness of machine learning models. Since then, more datasets have been introduced, and starting in 2016, a new dataset was released almost every year. All these datasets differ in the total number of reviews, subdomain, average review size, and percentage of included studies. However, the overall tendency shows a very high-class imbalance towards the negative class (i.e., irrelevant publications). Datasets introduced by Parmar [63] and Miwa et al. [54] are not publicly available, yet they were used in two and three research papers, respectively, so we included them in our comparison.

Until 2017 all of the datasets contained only the citation list with eligibility decisions [56]. More recently, datasets started to include titles of SLRs and search queries used for finding publications. Additional metadata is limited to search queries [71], review protocols (three datasets released as a part of the CLEF TAR shared-task by Kanoulas et al. [30, 31, 32]), review updates [2] and seed studies [88]. However, none of the datasets includes the eligibility criteria, the most critical section of SLR text used by manual annotators when assessing the relevance of publications. They also do not contain the information about why a particular paper was excluded from the review. Without this data, the automated citation screening problem cannot be tackled in any other way than a binary decision. This is not the case in real life, as a typical SLR contains at least several exclusion and inclusion criteria, and the decision about every paper can be presented as a multi-dimensional relevance problem.

So far, there has been little attention to review automation outside of the medical domain. The only available datasets are four social science reviews by Miwa et al. [54], and seven computer science

[9] Between the submission of the main paper and the supplementary materials, one more new citation screening dataset with 10 SLRs was released on 5 June 2023 [5].

reviews by Hannousse and Yahiouche [22]. Compared to the general interest and rate of production of SLRs in other domains, this overall underrepresentation of benchmark datasets could be improved. We also found one dataset containing one large SLR of environmental policies [26], which has a different scope and format than other datasets, so we decided not to include it in CSMeD yet.

Papers from the ML and NLP domains, very often evaluate their approaches on datasets introduced by Cohen et al. [11], which is, at the moment of writing this review, 17 years old. On the other hand, IR focused papers present their evaluation on CLEF TAR task datasets.

In terms of evaluation of classification approaches, aside from Precision and Recall, metrics include variations of the harmonised mean between the two, i.e. F$_\beta$–score, $utility$, $U19$ [82, 81, 83], sensitivity-maximising thresholds [13], and AUC [12]. Work Saved over Sampling (WSS) was proposed as a custom metric for evaluating this task as it measures the amount of work saved when using machine learning models to screen irrelevant publications [11, 52, 37, 38]. The True Negative Rate (TNR) was proposed as an alternative as it addresses some of the limitations of WSS regarding averaging scores from multiple datasets [40]. The measures of normalised Precision at r% recall (nPrecision@r%) and normalised rectified TNR at r% recall (nReTNR@r%) have also been introduced to focus on other important aspects of screening task: screening full texts and estimating users' time savings when compared to the random ranking, respectively [41].

Cost-based and economic-based metrics were also used, especially in the context of the query formulation task in the CLEF TAR shared task [32, 30, 31], e.g., total cost (TC) or total cost with a weighted penalty (TCW). The TREC Total Recall track [19] also used a cut-off based metric, $recall@aR + b$, which is defined as the recall achieved when $aR + b$ documents have been identified, where R is the number of relevant documents in the collection and a and b are parameters. When $a = 1$ and $b = 0$, $recall@aR + b$ is equivalent to R-precision. Finally, there has been a proposal to shift away from measuring Recall and instead evaluate how accurately automated methods can replicate the original systematic review outcomes [43].

The practical relevance of evaluating CS with metrics like the area under the ROC curve (AUC) [44] has been called into question, as it may not align with the goals of the citation screening task. Given that the CS task is primarily focused on achieving high recall, using AUC as an evaluation metric can be misleading, as it may highlight model improvements at lower recall values [40]. Having a unified benchmarking approach would also help to resolve these problems.

Finally, we were interested in checking how recently each dataset was used, where that usage was published, and what kind of evaluation measures were applied to that data. Table 7 presents the summary of our findings. We can see that to this date, most datasets were used in the past two years and simultaneously used by different publications. There is also a disparity in used evaluation measures, yet the basic Precision, Recall and F1-score prevail.

Table 7: Usage statistics of the SLR datasets, including the latest publication year, venue and evaluation measure. We report two usages in case there was a more recent pre-print published.

	Release - last time used	Evaluation schema (latest)	Venue (latest)
(1)	2006 - 2023 [45, 40]	TNR [40], AUC [45]	ECIR
(2)	2010 - 2022 [38]	WSS, Precision@95% [38]	ECIR
(3)	2014 - 2016 [23]	Yield, Burden, WSS [23]	JBI
(4)	2016 - 2022 [38], 2023 [44]	WSS, Precision@95% [38], AUC [44]	ECIR
(5)	2017 - 2018 [70]	Precision, Recall, WSS [70]	SIGIR
(6)	2017 - 2023 [89]	Precision, F1, Recall [89]	WSDM
(7)	2018 - 2023 [89]	Precision, F1, Recall [89]	WSDM
(8)	2019 - 2022 [87], 2023 [44]	MAP, Precision, nDCG [87], AUC [44]	ECIR
(9)	2019 - 2020 [3]	Recall, Precision [3]	JAMIA
(10)	2021 - 2022 [62]	F1-Score [62]	NAACL
(11)	2022 - 2023 [90]	Precision, F1, F3, Recall [90]	SIGIR
(12)	2022 - 2022 [22]	Recall, Precision, Macro F1, Accuracy [22]	MedPRAI

A.2 SLR datasets in biomedical benchmarks

Systematic literature reviews consist of multiple steps, and depending on the granularity, previous studies enumerated between four and up to 15 tasks that might be included in the SLR process [76]. High-level tasks include steps of preparation, followed by the search and appraisal of primary studies and then synthesis and write-up of the evidence. According to van Dinter et al. [78], citation screening (selection of primary studies) was the step for which most of the automation-related research was happening. Among other steps, the tasks of query formulation, information extraction, risk of bias assessment, and, more recently, text summarisation were also introduced.

Marshall et al. [51] introduced a large dataset with Cochrane reviews for the task of assessing the risk of bias – a procedure aiming at establishing the quality of input studies. Nye et al. [59] proposed a PICO (Population, Intervention, Comparison and Outcome) extraction dataset containing 5,000 annotated abstracts of biomedical publications. In the query formulation, often the models evaluate their performance on the CLEF TAR 2017-2018 datasets [30, 31]. For the task of systematic review summarisation, a shared task was introduced [86] consisting of two datasets: [84, 14].

In a comprehensive catalogue of medical artificial intelligence datasets and benchmarks by Blagec et al. [8], only three citation screening datasets are mentioned: Cohen et al. [11], Wallace et al. [81], and Miwa et al. [54]. Of these three datasets, only two are publicly available, and both are already implemented in CSMED. Additionally, another five private SLRs used in only one publication [73] are mentioned.

There is poor coverage of SLR datasets among biomedical benchmarks, especially for the task of citation screening. None of the existing benchmarks contains any publicly available citation screening dataset. Only the BoX [62] benchmark uses five SLRs, but these datasets are private and cannot be obtained even through a DUA (Data Use Agreement).

From other SLR automation tasks, BigBio [16] and BLURB [20] benchmarks contain only one information extraction dataset by Nye et al. [59]. BLUE [64] and CBLUE [96] benchmarks do not contain any SLRs-related task. Therefore, there is a clear need to develop and include publicly available SLR datasets in biomedical benchmarks, particularly for citation screening tasks, to facilitate further research and progress in this field.

The latest advances in Large language models (LLMs) offer significant potential for aiding in SLR automation but simultaneously raise several concerns. A user study by Yun et al. [94] mentions that SLR practitioners acknowledged the potential utility of LLMs in various tasks, such as generating the first draft of a review, writing plain language summaries, and extracting information from longer texts. On the other hand, domain experts have highlighted several crucial issues, including concerns about hallucinations, the untraceable origins of generated content, and the proliferation of bad-quality reviews.

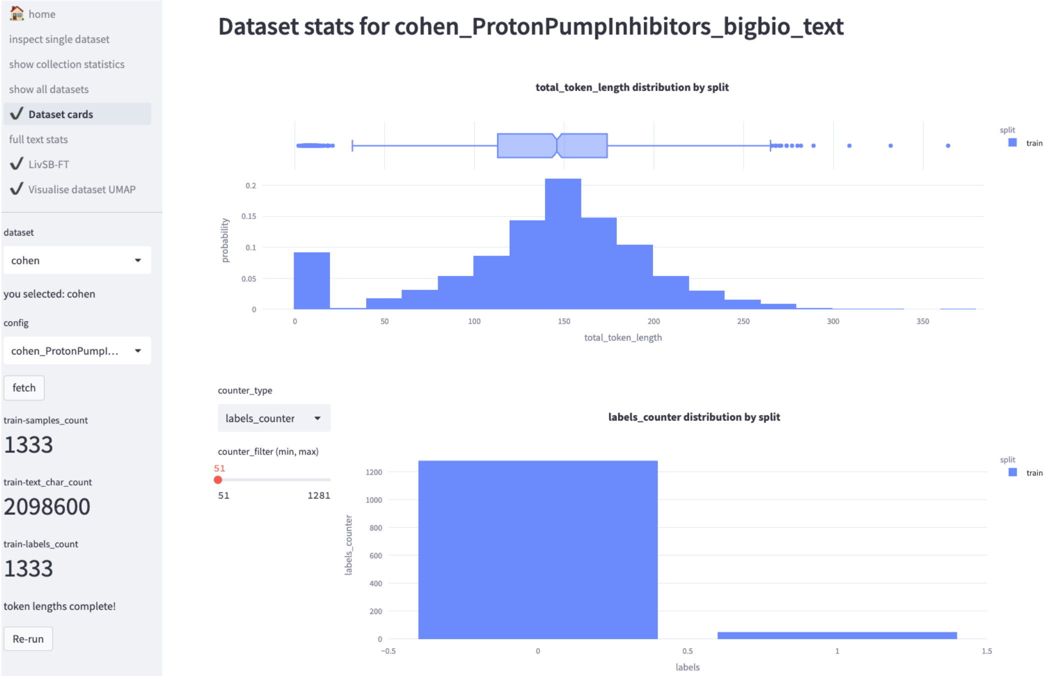
B Visualisations

We leverage Streamlit[10] to create interactive visualisations for our meta-dataset. We present essential details for every dataset, such as the number of training samples, character and word counts, and labels and token lengths distribution across dataset splits (example in Figure 2). We build upon the existing BigBio schemas and visualisations, extending them to incorporate citation screening-specific details. We also build a dedicated page to explore CSMED-FT dataset containing full text documents.

We further focus on measuring the overlap between datasets. We check for the overlap on the level of systematic reviews based on the review's Cochrane ID. This can help researchers understand potential biases, redundancy, or complementary aspects across various datasets.

We use a TF-IDF-based document vectoriser with UMAP [53] to plot two-dimensional representations of the datasets. This approach allows us to effectively capture and display the structural patterns and similarities within a single systematic literature review, aiding researchers in identifying clusters, outliers, and potential data correlations. An example of UMAP clustering of publications is presented in Figure 3.

[10] `https://streamlit.io`



Dataset stats for cohen_ProtonPumpInhibitors_bigbio_text

dataset: cohen

you selected: cohen

config: cohen_ProtonPumpl...

train-samples_count
1333

train-text_char_count
2098600

train-labels_count
1333

token lengths complete!

counter_type: labels_counter

counter_filter (min, max): 51 – 51 1281

Figure 2: Example visualidation with statistics for a “Proton Pump Inhibitors” SLR dataset.

A live demo of the visualisation interface is available under the following URL: `https://systematic-review-datasets.streamlit.app/`. Some features require data preprocessing; they are unavailable in the demo but can be run locally using the code from the GitHub repository.

C CSMeD data card

Dataset Description: CSMeD is a meta-dataset consisting of nine different citation screening datasets containing 325 systematic literature reviews (SLRs). Each systematic review consists of a list of publications that need to be classified as either *relevant* or *irrelevant*. All datasets have data loader scripts providing programmatic access aligned with the BigBio framework and HuggingFace datasets library. We preserve the original splits of the datasets. We also generate data cards for every dataset which is part of the CSMeD. CSMeD allows for accessing independent datasets and single systematic reviews, which are part of each dataset.

TRAIN-COCHRANE and DEV-COCHRANE splits contain expanded metadata about systematic reviews such as systematic review title, abstract, eligibility criteria and search strategy. TRAIN-BASIC is a set of SLRs for which such meta-data was unavailable and it is characterised by the systematic literature review title. TRAIN-COCHRANE and DEV-COCHRANE splits are suitable for the tasks of question answering, natural language inference, and text pair classification. TRAIN-BASIC is suitable only for the text classification task.

Homepage: `https://github.com/WojciechKusa/systematic-review-datasets`

URL: `https://github.com/WojciechKusa/systematic-review-datasets`

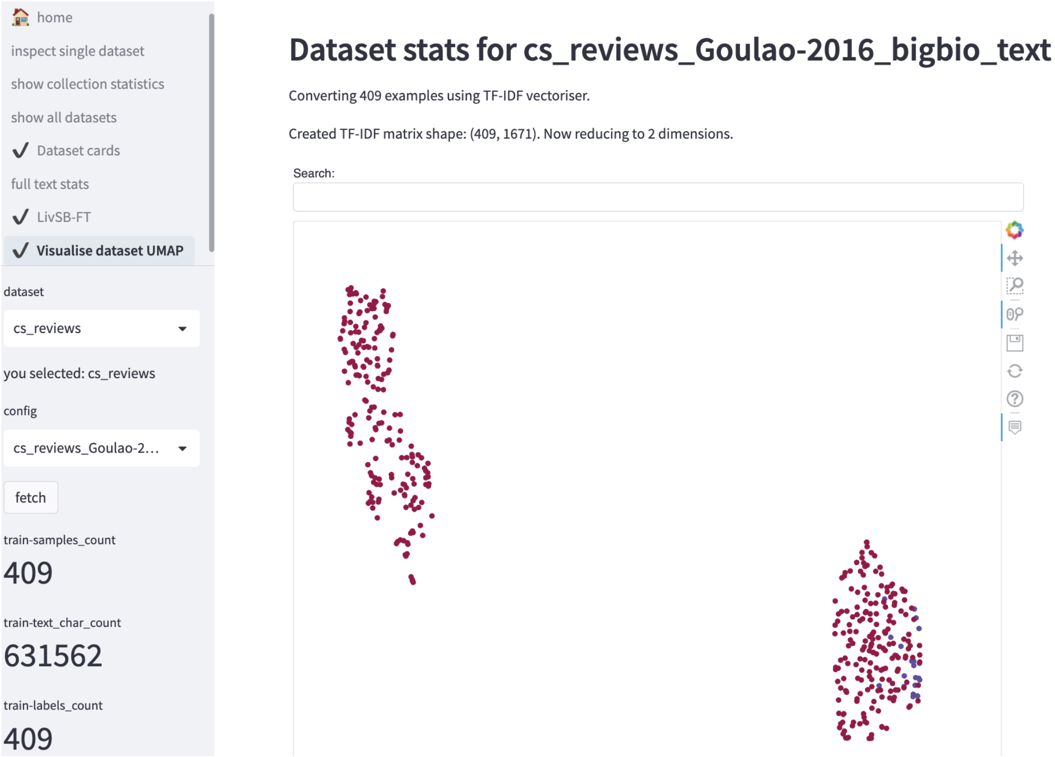
Licensing: CC BY 4.0

Languages: English

Tasks: text classification (`TXTCLASS`), question answering (`QA`), natural language inference (`NLI`), text pairs classification (`PAIRS`).

Schemas: Text (`TEXT`), Text pairs classification (`PAIRS`). Question Answering (`QA`), source (`source`).

Splits: TRAIN-BASIC, TRAIN-COCHRANE, CSMeD-DEV-COCHRANE, CSMeD-ALL



Dataset stats for cs_reviews_Goulao-2016_bigbio_text

Converting 409 examples using TF-IDF vectoriser.

Created TF-IDF matrix shape: (409, 1671). Now reducing to 2 dimensions.

Search:

dataset: cs_reviews

you selected: cs_reviews

config: cs_reviews_Goulao-2...

train-samples_count: 409

train-text_char_count: 631562

train-labels_count: 409

Figure 3: Example visualisations with TF-IDF and UMAP representation of documents for a "CS-Goulao-2016" SLR. Based on the plot, one can see that the retrieved documents are grouped in two clusters with all relevant publications belonging to one of them (bottom-right part of the plot). This can be an indicator that any model will likely remove the other "non-relevant" cluster of documents and hence achieve good score in detecting true negatives.

D CSMED-FT

CSMED-FT is an extension of the CSMED meta-dataset that specifically focuses on the full text screening step in SLRs. CSMED-FT is to the best of our knowledge, it is the first dataset explicitly targeted at the screening of the full text of publication. While previously researchers already used full text screening labels from other datasets to evaluate their models, the input to these models constituted only the titles and abstracts of publications [28].

D.1 Dataset construction details

To construct CSMED-FT, we collected various elements of SLRs from the Cochrane Library website, including the title, abstract and eligibility criteria sections of the SLR and SLRs' appendix and references. The appendix contains a search strategy, while the references list papers categorised as: "studies included in the review", "studies excluded from the review", and "additional references". We decided to focus solely on the "included" and "excluded" categories as there is no definitive way to determine the intended meaning when researchers added papers as additional references. However, in future work, we plan to explore the possibility of extending the dataset to encompass publications from the "additional references" category.

To obtain the full texts of references, we used the DOI (Digital Object Identifier) of each publication. While some references directly provided the DOI, for others, we initially attempted to match them to PubMed IDs and then extracted the DOIs from PubMed and Semantic Scholar. To assign PubMed IDs to the publications parsed from the Cochrane website, we followed a four-step process:

- We check if the PubMed ID information is provided on the Cochrane references webpage.

- We conduct search in PubMed using Entrez[11] by searching for the same title and authors.
- We search for the PubMed ID in SemanticScholar using publication DOI from Cochrane references webpage.
- We search again in PubMed, this time with a relaxed requirement by searching for an exact match in the title only.

We then use the PubMed ID to resolve the DOI of the publication. We could match the DOI for more than 61% of references.

We adopted a time-wise construction approach for CSMeD-FT canonical splits to ensure the integrity and avoid data contamination. Therefore, we selected 29 SLRs not part of any previously released datasets to form our test set. We used data from previous publications to construct a testing and development set: dataset used by Nussbaumer-Streit et al. [57] for the development set and dataset introduced by Scells et al. [71] for training split. It should be noted that newer SLRs tend to have more comprehensive metadata and more open-access full text publications available. This resulted in token length and label frequency differences across the dataset splits (Figure 4). Despite these variations, we decided to retain these splits as they present a more realistic and challenging scenario, closely reflecting real-life circumstances.

We have made the entire dataset construction procedure available in our repository, enabling transparency and reproducibility.

D.2 CSMeD-FT Data Card

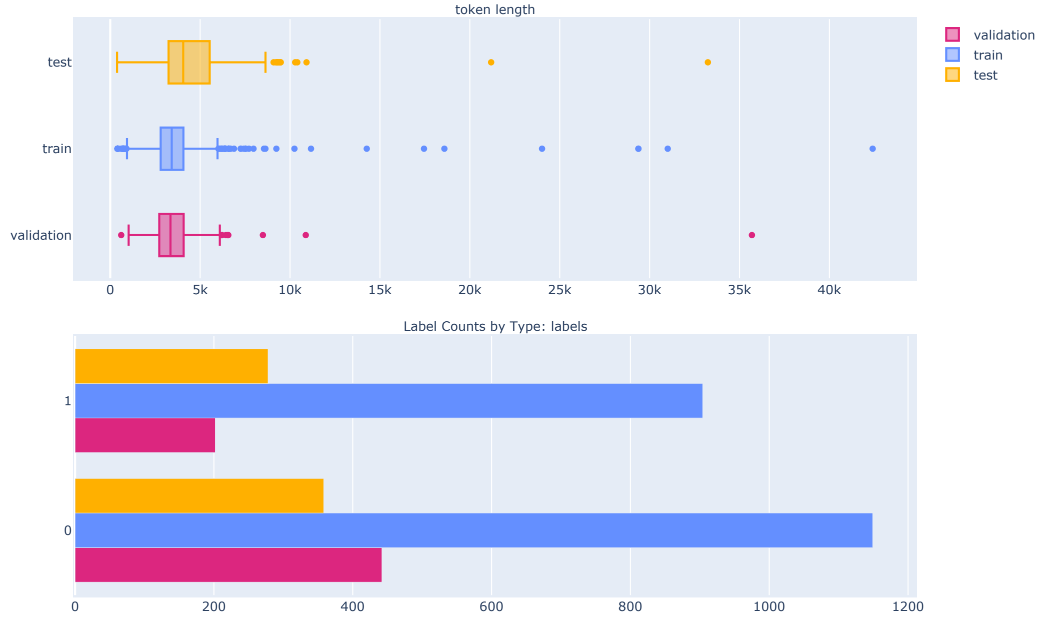


Figure 4: Token frequency distribution by split (top) and frequency of different kind of instances (bottom).

Dataset Description The dataset focuses on the task of full text screening for systematic literature review creation. It contains 3,333 systematic literature review and publication pairs with decisions if the publication was included in the systematic literature review. Every excluded publication also contains a textual explanation of why it was excluded. Systematic literature reviews are formatted in a JSON format, whereas publications are stored as CSV files. CSMeD-FT-SAMPLE is a subset of CSMeD-FT-TEST dataset. We intend to store the dataset on the TU Wien Research Data repository,[12] currently the dataset is available on the project GitHub repository.

[11] https://www.ncbi.nlm.nih.gov/search/

[12] https://researchdata.tuwien.ac.at

Homepage: `https://github.com/WojciechKusa/systematic-review-datasets`

URL: `https://github.com/WojciechKusa/systematic-review-datasets/raw/main/data/CSMeD/CSMeD-FT.zip`

Licensing: CC BY 4.0

Languages: English

Tasks: text pairs classification, natural language entailment

Schemas: `TEXT`, `PAIRS`, source.

Splits: TRAIN, DEV, TEST, SAMPLE

Dataset size (document pairs): TRAIN: 2,053, DEV: 644, TEST: 636, SAMPLE: 50

Size of downloaded dataset files: 33.5 MB

Size of the generated dataset files: 112.2 MB

E Examining dataset overlap

We evaluate the overlap between datasets at the level of entire systematic reviews. This analysis aims to understand the potential duplication of information and data leakage across different datasets.

Table 8 presents the extent of overlap observed between the train and test splits of the datasets. The TAR 2019 collection is most severely affected, with 3 SLRs duplicated in its train and test splits. SLRs released as part of the SIGIR 2017 collection [71] are also present among the test splits in CLEF TAR 2017 and 2019 collections.

Table 8: List of overlapping Cochrane systematic literature reviews between datasets.

Cochrane review ID	First collection	Other collections
CD011145	sigir2017 (train)	tar2017 (test)
CD010633	sigir2017 (train)	tar2017 (test), tar2018 (train), tar2019 (train)
CD010653	sigir2017 (train)	tar2017 (test), tar2018 (train), tar2019 (train)
CD010542	sigir2017 (train)	tar2017 (test), tar2018 (train), tar2019 (train)
CD009185	sigir2017 (train)	tar2017 (test), tar2018 (train), tar2019 (train)
CD008081	sigir2017 (train)	tar2017 (test), tar2018 (train), tar2019 (train)
CD002143	sigir2017 (train)	sigir2017 (train)
CD001261	sigir2017 (train)	tar2019 (test)
CD011571	tar2019 (train)	tar2019 (test)
CD012164	tar2019 (train)	tar2019 (test)
CD011686	tar2019 (train)	tar2019 (test)

It is worth noting that we did not explicitly report the overlap between different CLEF TAR datasets [30, 31, 32]. The owners of the dataset have already acknowledged that each new edition of the dataset includes SLRs from the previous editions as part of the training data. As the older datasets did not share metadata about the considered reviews (except for the very high-level title of the review (e.g. ADHD or COPD), we did not have access to the mapping to the published reviews.

F Zero-shot screening on CSMeD

Our proposed approach is based on the recent advancements in language modelling to conduct a zero-shot ranking or classification of papers. We consider metadata in the CSMeD-COCHRANE dataset as various query representations for measuring their impact on screening.

In this experiment, we evaluate the impact of the SLR protocol section on ranking for statistical and neural models in a zero-shot setting. We use two statistical models BM25 and TF-IDF, and

Table 9: Results of zero-shot evaluation on CSMeD-cochrane-dev dataset. For each measure, **bold** values indicate the highest score for each model across query representation. Underlined values indicate the highest score across all tested models.

Model	Representation	TNR@95%	nP@95%	Last Rel	nDCG@10	MAP	R@10	R@50	R@100
BM25	Title	0.469	0.142	72.2	0.438	0.388	0.349	0.623	0.704
	Abstract	**0.474**	**0.170**	**63.6**	**0.503**	**0.453**	**0.379**	**0.657**	**0.757**
	Search strategy	0.379	0.093	72.1	0.336	0.311	0.268	0.507	0.625
	Criteria	0.430	0.145	67.0	0.452	0.417	0.345	0.629	0.725
TF-IDF	Title	0.439	0.126	75.1	0.334	0.322	0.295	0.575	0.661
	Abstract	**0.490**	**0.147**	**62.8**	**0.417**	**0.404**	**0.348**	**0.640**	**0.728**
	Search strategy	0.372	0.078	72.9	0.271	0.272	0.233	0.500	0.595
	Criteria	0.453	0.139	67.0	0.375	0.372	0.305	0.616	0.704
MiniLM	Title	0.472	0.217	68.1	0.470	0.414	0.379	0.673	0.763
	Abstract	0.492	**0.240**	65.5	**0.517**	0.451	**0.398**	**0.680**	**0.782**
	Search strategy	0.411	0.171	71.4	0.370	0.346	0.320	0.609	0.688
	Criteria	**0.527**	0.198	**60.9**	0.497	**0.456**	0.384	0.657	0.747
MPNet	Title	0.467	0.230	66.6	0.476	0.429	0.376	0.684	0.774
	Abstract	0.516	**0.265**	63.8	**0.556**	0.482	**0.420**	**0.692**	0.777
	Search strategy	0.429	0.181	68.6	0.400	0.372	0.328	0.614	0.699
	Criteria	**0.545**	0.216	**58.5**	0.514	**0.488**	0.393	0.691	**0.784**
BioBERT	Title	0.439	0.141	66.7	0.391	0.369	0.337	0.624	0.717
	Abstract	0.494	0.166	64.4	0.463	0.448	**0.367**	0.655	**0.768**
	Search strategy	0.369	0.098	72.9	0.350	0.335	0.273	0.523	0.635
	Criteria	**0.507**	**0.182**	**62.7**	**0.494**	**0.468**	0.358	**0.681**	0.765

three Transformer-based models: MiniLM-L6-v2[13], mpnet-base-v2[14] and BioBert-snli[15] from the SentenceTransformers library [65]. MiniLM model uses 256 tokens, whereas MPNet and BioBERT use 512 tokens.

We test four different SLR meta-data sections from the SLR protocol as input representations: (1) title, (2) abstract, (3) search strategy and (4) eligibility criteria. Predictions are run on the CSMeD-cochrane-dev split. We use the `retriv` Python library for implementing the pipeline [6]. The code and detailed instructions for replicating our results are available at `https://github.com/WojciechKusa/CSMeD-baselines`.

F.1 Evaluation

We select True Negative Rate at 95% Recall ($TNR@95\%$) and normalised Precision at 95% Recall ($nP@95\%$) as primary evaluation measures. We also evaluate the average position at which the last relevant item is found [30, 31, 32], calculated as a percentage of the dataset size ($Last\ Rel$). Lower values of $Last\ Rel$ indicate better performance. Additionally, we compute traditional evaluation measures: $nDCG@10$, MAP and Recall at rank k ($R@k$), with k in {10, 50, 100} following the evaluation from Kanoulas et al. [30].

F.2 Expanded results

Table 9 presents complete results on the CSMeD-dev-cochrane dataset. Overall, we find that models using SLR abstracts and eligibility criteria perform the best with the consistent superiority of neural network-based models over traditional retrieval models. The topical similarity between the publications and the SLR abstract suggests an important role of the abstract in the automated screening process.

Across all measures for both statistical models, representing SLR using its abstract consistently outperforms others. This indicates that abstracts, as a source of external knowledge, contain more comprehensive and relevant information for automated citation screening compared to titles or search strategies. This finding is aligned with the analysis of the statistical models for the clinical

[13] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

[14] `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`

[15] `https://huggingface.co/pritamdeka/S-BioBert-snli-multinli-stsb`

Table 10: Systematic literature review protocol section lengths in number of words for CSMeD-cochrane-all dataset.

Word count	**Abstracts**	**Titles**	**Search strategy**	**Eligibility criteria**
Mean	720.8	10.8	567.6	852.2
25th Percentile	574.0	7.0	129.5	450.5
50th Percentile	718.0	10.0	273.0	662.0
75th Percentile	878.0	13.0	610.0	1005.0
90th Percentile	976.0	17.0	1196.8	1503.4

trials matching task, which also showed the inability of these models to comprehend the eligibility criteria [42].

On the other hand, more advanced neural models tend to utilise the eligibility criteria information better. $TNR@95\%$ is higher when using the criteria information for all three considered Transformer-based models. Similar considerations can be given about other evaluation measures, where we notice that with growing model size and input window, their performance is getting better when using the criteria section compared to SLR abstract. However, It should be noted that the criteria section is typically more relevant to the full text screening step than title and abstract screening.

The best-performing model, MPNet, using SLR eligibility criteria, achieves $TNR@95\%$ equal to 0.545, meaning that this model can remove, on average, more than half of the true negatives when achieving a recall of 95%. We also see that TNR and nP measures are not always aligned between model and representation combination.

Table 10 shows the word count statistics of SLR sections for CSMeD-cochrane-all dataset. The text is truncated for more than half of the examples in the case of SLR abstracts and eligibility criteria. This also prevents the use of the cross-encoder approach, where the concatenated publication and SLR section would exceed the maximum context window for typical BERT-style models. Using models allowing for longer input sequences could enhance the ranking quality. Exploring large language models or advanced training scenarios like the Topical-Criteria Re-Ranking curriculum learning [42] might also reveal the potential for further improving the results.

G CSMeD-ft experiment setup

The code and detailed instructions for replicating our results are available at `https://github.com/WojciechKusa/CSMeD-baselines`.

G.1 Transformer model fine-tuning

We select the following model checkpoints from HuggingFace Transformers library:

- Longformer-base – `https://huggingface.co/allenai/longformer-base-4096`
- BigBird-roberta-base – `https://huggingface.co/google/bigbird-roberta-base`
- Clinical-Longformer – `https://huggingface.co/yikuan8/Clinical-Longformer`
- Clinical-BigBird – `https://huggingface.co/yikuan8/Clinical-BigBird`

We want to decide whether a publication fulfils all inclusion criteria and none of the exclusion criteria to include it in the SLR. Specifically, this means matching the eligibility criteria of SLR with the full text of the candidate publication. As input, the model receives the text of the review and publication and is asked to predict a binary category. We concatenate the review title with the eligibility criteria section to create the review text. For publications, we concatenate the title, abstract and the main text.

As available input text (review text + publication text) almost always exceeds the available context window of considered models (4,096 tokens), we use the following approach to allocate available space. We use the `TokenTextSplitter` method from the langchain library[16] with the gpt-3.5-turbo-

[16] `https://github.com/hwchase17/langchain`

0301 model to select the review text that would fit the context window. We select at most half of the available context window, so in the context of all Transformer models, review text equals at most 2,048 tokens. This action truncates some part of the eligibility criteria section, i.e. for 13% of items in the trainset and 42% in the test set (Table 11). We fill the remaining input sequence with the publication text.

Table 11: Statistics of a review text with respect to the fit within 2,048 tokens context window.

	CSMeD-FT-TRAIN	CSMeD-FT-DEV	CSMeD-FT-TEST	CSMeD-FT-SAMPLE
Avg # splits	1.13	1.24	1.83	1.74
Median # splits	1	1	1	1
Max # splits	2	2	4	4
Min # splits	1	1	1	1
More than 1 splits	13%	24%	42%	42%

We run our experiments on a single server with 4 Nvidia RTX 3090 GPUs with 24GB of RAM each. We use a per-device batch size of 1 with eight gradient accumulation steps. We test several learning rates with the best results for 1e-5, and we set weight decay to 0.01. We use AdamW [50] with default values of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We evaluate models after each epoch on the validation set and select the model with the highest macro F1-score.

One training epoch took around 30 minutes both for BigBird and Longformer-based models. For inference, Longformer architecture processed, on average, 2.9 samples per second, whereas BigBird models 2.65 samples per second. Making predictions on the entire test split of 636 documents took less than 4 minutes for all models.

G.2 Zero-shot language model evaluation

Similarly, as for the fine-tuned classification models, we reserve at most half of the context window size for the systematic literature review description and fill the remaining tokens with the publication text. We measure the text length using the OpenAI library tiktoken[17], which provides tokenisers for GPT-3.5 and GPT-4 models. We use the `openai` python library version `0.27.7`, and use the default chat completion function parameters of temperature = 1 and top_p = 1.

We set our total budget to 50 USD and conduct the experiments only on the CSMeD-FT-TEST-SMALL subset for GPT-4 model. For the GPT-3.5-turbo-16k model, making predictions on all 636 examples of the CSMeD-FT-TEST split took 44 minutes. However, this value was heavily influenced by the default OpenAI's rate limits of 180,000 tokens per minute for our organisation. We use the following prompt template:

Input Template:

```
Does the following scientific paper fulfill all eligibility criteria and \
should it be included in the systematic review? \
Answer ‘Included’ or ‘Excluded’. \
Systematic review: "{{r.title}}" \n "{{r.criteria}}" \n\n \
Publication: "{{p.title}}" \n "{{p.abstract}}" \n "{{p.main_text}} \n\n \
Answer:
```

Output Template:

```
{{label}}
```

Answer Choices:

```
Included ||| Excluded
```

[17] https://github.com/openai/tiktoken