Extreme Systematic Reviews: A Large Literature Screening Dataset to Support Environmental Policymaking

Jingwen Hou*
University of Pennsylvania
jwhou@seas.upenn.edu

Xiaochen Wang*
The Pennsylvania State University
xmw5190@psu.edu

Jean-Jacques Dubois
United States Environmental Protection Agency
dubois.jeanjacques@epa.gov

R. Byron Rice
United States Environmental Protection Agency
rice.byron@epa.gov

Amanda Haddock
United States Environmental Protection Agency
haddock.amanda@epa.gov

Yue Wang
The University of North Carolina at Chapel Hill
wangyue@unc.edu

ABSTRACT

The United States Environmental Protection Agency (EPA) periodically releases Integrated Science Assessments (ISAs) that synthesize the latest research on each of six air pollutants to inform environmental policymaking. To guarantee the best possible coverage of relevant literature, EPA scientists spend months manually screening hundreds of thousands of references to identify a small proportion to be cited in an ISA. The challenge of extreme scale and the pursuit of maximum recall calls for effective machine-assisted approaches to reducing the time and effort required by the screening process. This work introduces the ISA literature screening dataset and the associated research challenges to the information and knowledge management community. Our pilot experiments show that combining multiple approaches in tackling this challenge is both promising and necessary. The dataset is available at https://catalog.data.gov/dataset/isa-literature-screening-dataset-v-1.[1]

CCS CONCEPTS

• **Information systems** → **Document filtering**; • **Computing methodologies** → **Supervised learning by classification**.

KEYWORDS

High-Recall Retrieval, Policy Relevance, Systematic Review

ACM Reference Format:
Jingwen Hou, Xiaochen Wang, Jean-Jacques Dubois, R. Byron Rice, Amanda Haddock, and Yue Wang. 2022. Extreme Systematic Reviews: A Large Literature Screening Dataset to Support Environmental Policymaking. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3511808.3557600

*Jingwen Hou and Xiaochen Wang contributed equally to this research.
[1]The views expressed in this article are those of the author(s) and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557600

1 INTRODUCTION

Since 1977, the U.S. Clean Air Act (CAA) has directed the United States Environmental Protection Agency (EPA) to synthesize the most recent scientific research at regular intervals for each of six criteria ambient air pollutants. These documents, called Integrated Science Assessments (ISAs), aim to provide an updated, comprehensive review of the state of the science on the health and welfare effects of these air pollutants, and lay the scientific foundation for the policymakers setting the National Ambient Air Quality Standards (NAAQS) [2]. Because an ISA document has far-reaching implications on environmental policies and subsequent impacts on public health and welfare, the EPA invests substantial resources to ensure that the document covers up-to-date policy-relevant literature as comprehensively as possible. As a result, ISAs can be viewed as extremely large scale, periodically updated systematic literature reviews that present unique characteristics and challenges as described below.

Large scale: To give a concrete sense of scale, the 2020 ISA for Ozone and Related Photochemical Oxidants contains 12 chapters spanning 1,468 pages, and 1,704 unique citations [1]. Initial high-recall search queries returned a total of 171,376 potentially relevant publications, which were reviewed by a team of 38 domain experts for more than 6 months.

High-recall requirement: Missing any relevant reference may have substantial consequences for public health and ecosystem welfare. In the literature screening process, the EPA scientists pursue maximum recall despite enormous cost of time and effort. To guarantee the best possible coverage of relevant literature, an ISA is critically reviewed and validated through as many as three drafts by a committee of domain experts (the Clean Air Scientific Advisory Committee), by interest groups and by the public, who recommend additional references for citation should any have been missed.

Periodicity: This labor-intensive screening and validation processes are conducted at regular intervals for each of the six air pollutants to meet the CAA's five-year requirement. Every iteration is based on the previous one. Given the growth of publications, the effort has increased substantially with every iteration.

Topical breadth: Each ISA covers not only multiple scientific disciplines, including epidemiology, toxicology, ecology and more, but also an extreme diversity of study types. An ISA is not a simple concatenation of these disciplines and studies, but tightly integrates them into an organic whole.

Policy relevance: Because the goal of an ISA is to inform policymaking, cited literature need to be both topically relevant and *policy relevant*. Among all the articles about a pollutant and its effects, only those with the potential to inform risk assessment and the setting of NAAQS directly or indirectly are policy relevant.

All these characteristics clearly distinguish the ISA literature screening task from those in typical systematic reviews [10]. Commonly seen in evidence-based medicine, systematic reviews usually focus on a sharply circumscribed topic with an orders-of-magnitude smaller workload and data scale. For example, they often screen hundreds to thousands of search results to find a few dozen relevant ones to be referenced and reviewed [3, 13, 16, 18].

The EPA has had some success using commercial implementations of machine-assisted review technologies that employ active learning, but they have not reduced the time and labor spent on reviewing literature that will not be cited in the ISA as drastically as needed. This paper presents an ISA literature screening dataset both as a resource and as a "call to action" that invites research efforts on the associated challenges.

Relevance and Expected Impact. This data resource is relevant to several fields in the information and knowledge management community. First, it introduces *policy relevance*, a new dimension of relevance of interest to the field of information retrieval. Second, the high-recall ranking problem is a testbed for machine learning techniques. We show in Section 3 that it can be approached from supervised learning, active learning, and transfer learning perspectives. Third, this can also be seen as a content recommendation problem when the collection of screened literature is recast as a continuous data stream over multiple years. Fourth, the data is of interest to the scientific literature management community as an extreme case of machine-assisted approaches for systematic review. Research on this dataset can have far-reaching real-world impact. Regulatory scientists need more efficient tools and methods to gather policy-relevant literature. These tools and methods shape the scientific foundation for environmental policies, which directly impact public health and ecological welfare.

2 ISA LITERATURE SCREENING DATASET

We provide six data files, three for each of two successive ISAs on ozone (2013, 2020): reference metadata, citation context, and semantic map. Below we explain their generating processes.

2.1 Reference Metadata

During the literature screening process, EPA scientists use two strategies to search for articles. The first strategy uses high-coverage Boolean strings of topical keywords to search various databases, including PubMed and Web of Science. The second strategy is citation relational search, where references from the previous ISA are used as "seeds" and any article that has cited any of the seeds since the release of the previous ISA is returned. These searches are conducted for each chapter to cover different aspects of research. The results are then aggregated and deduplicated, giving rise to the **search result set**, which we denote as $\boldsymbol{S}$.

The set of all references cited in the ISA is a mixture of the more important policy-relevant references that require high recall, and supplementary references that provide summaries, complete the narrative of legislative and historical background, the explanation of research instruments and methodology, and the development process of the ISA itself. ISAs are structured such that all policy-relevant references are contained in core sections that scientists can readily identify (see Section 2.3), where the detailed reviewing of evidence takes place. References cited in the core sections are defined here as the **core reference set**, which we denote as $\boldsymbol{C}$. The intersection of the core references and the search results are defined as the **relevant reference set**, which we denote as $\boldsymbol{R}$. Note that the search result set S may not fully cover the core reference set C because some references in the core sections may still be supplementary and added outside of the systematic search process.

For each ISA, we release the reference metadata for all articles in $C \cup S$. The columns include unique identifiers, authors, year, title, abstract, and indicators of whether an article is in C or S.

	Ozone 2013	Ozone 2020
# of search results, $\|S\|$	15,772	171,376
# of core references, $\|C\|$	2,153	1,349
# of relevant references, $\|R\|$	2,063	1,153
avg. # of words/title	13.57	15.23
avg. # of words/abstract	246.17	239.14
# of citation instances	5,949	3,887
# of sections	249	289
# of distinct context paragraphs	1,836	1,405
avg. # of words/paragraph	112.86	79.36
# of chapter categories	5	5
# of topics	43	43
# of disciplines	7	7
avg. # of sections/chapter category	77.00	120.20
avg. # of sections/topic	6.49	7.14
avg. # of sections/discipline	39.29	49.71

Table 1: Basic statistics for 2013-2020 ozone ISA datasets. The three parts correspond to statistics of three types of data: reference metadata, citation context, and semantic map.

2.2 Citation Context

Although the core references in C share the identity that they are "policy relevant", they are very far from semantically homogeneous given the breadth of topics and disciplines. Knowledge of the textual context in which the ISA cites references in C can be useful for modeling C in a fine-grained manner.

For each ISA, we provide the context in which a reference in C is cited. We parsed the ISA as a tree of document objects (section headings and associated paragraphs) and extracted the **sections** and the **context paragraphs** in which a reference is cited. The same reference can be cited in multiple sections and context paragraphs, each called a *citation instance*. We use *section* to broadly refer to any level of outline, ranging from Level-1 sections or chapters to Level-6 sections. We define the scope of a section as paragraphs between that section's heading and the next section's heading. Therefore, paragraphs in a child section do not belong to its parent section. In this way, all sections form a set-theoretic partition on all paragraphs (and hence all citation instances) in an ISA.

2.3 Semantic Map

As mentioned before, the EPA periodically updates ISAs on each of the six air pollutants. Although the outline or structure of an earlier ISA can differ from that of a later ISA, disciplines and core topics

remain largely stable over time. For example, in the 2013 ozone ISA, the atmospheric science chapter is "3 Atmospheric Chemistry and Ambient Concentrations", while in 2020, the chapter is "Appendix 1 Atmospheric Source, Chemistry, Meterology, Trends, and Background Ozone". Knowledge about how chapters and sections in different ISAs map to the same set of disciplines and topics can be useful in modeling persistent themes in the content (and therefore cited references) in ISAs over time.

For each ISA, we provide a semantic map from a section to its corresponding semantic labels. Three types of semantic labels are given. **Chapter category** describes a coarse-grained topic. All sections under the same chapter have the same chapter category. **Topic** assigns a fine-grained topic to each section. **Discipline** assigns one of seven disciplines to a section. Sections under the same topic can have different disciplines. The topic of background and summary sections are labeled as "Supplementary". Non-supplementary sections are core sections and references in those sections form the core reference set R. The semantic map does not contain an ISA's opening chapters that summarize later chapters (e.g., "Integrated Synthesis"), as those chapters are entirely supplementary in the perspective of this work.

Table 1 shows the basic statistics of the three data files for the two successive ozone ISAs. The topic "Supplementary" is excluded from "# of topics" and "avg. # of sections/topic". In the 2013 (2020) ozone ISA, 106 (298) sections are supplementary, respectively.

3 THE HIGH-RECALL RANKING PROBLEM

3.1 Problem Formulation

EPA scientists' most pressing need is to reduce manual efforts in reviewing hundreds of thousands of search results to identify policy-relevant references. It translates into a high-recall ranking problem: "to rank references in one ISA's S set such that *all* core references in the subset R are ranked high and identified early in the reviewing process." Given the richness of the above data resources, the problem can be formulated in various ways.

Supervised learning: Because an ISA on the same pollutant is updated iteratively, it is reasonable to assume that the notion of relevance from past ISAs generalizes to future versions. Thus the high-recall ranking problem can be approached from a supervised learning perspective. That is, a ranker can be trained on the data of the previous ISA and applied to the S set of the current ISA. This approach produces a static ranking of S.

Classical active learning: The problem can also be approached in an active learning framework [8, 17]. At the beginning of the reviewing process, scientists assign positive (relevant) and negative (non-relevant) labels to a small random subset of references, which can be used to train an initial machine learning classifier. In subsequent rounds, the classifier proactively selects informative references and asks for their labels. Those labels are then used to update the classifier. Unlike the static ranking produced by supervised learning, active learning can improve the classifier and hence the classifier-induced ranking of S continuously.

Active learning with knowledge transfer: The above two approaches can join force: one can transfer the knowledge learned from past ISA data to the active learning classifier. The benefit is that the transferred knowledge can save the labeling efforts of active learning, especially at the beginning of the process when no data is labeled, a problem known as the *cold start problem* [5].

Below we describe our preliminary experiments and results on the high-recall ranking problem. As the primary goal of this paper is to introduce the dataset and associated research opportunities, we intentionally use simple, standard algorithms to provide a preview and baseline references for future research.

3.2 Supervised Learning Experiments

Let us denote the 2013 (2020) ozone ISA search result set as S_1 (S_2), relevant reference set as $R_1 \subset S_1$ ($R_2 \subset S_2$), respectively. We learn a scoring function $s : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ denotes the feature space of references, S_1, S_2 (and hence R_1, R_2) are subsets of $\mathcal{X}$. $s(x)$ is trained on S_1, R_1 and applied to references in S_2. The hope is that $s(x)$ will assign higher scores to references in R_2 than those in $S_2 - R_2$.

3.2.1 Text-based Simple Ranker. We represent each reference by concatenating its title and abstract texts. The features are TFIDF-weighted bag-of-ngrams ($n = 1, 2, 3$). We train a logistic regression (LR) classifier on the training set $\{(x, y_t)|x \in S_1, y_t = \mathbf{1}\{x \in R_1\}\}$, where $\mathbf{1}\{z\} = 1$ if z is true and 0 otherwise. Throughout this paper, a LR classifier uses L_2 regularization with $C = 1$ and cost-sensitive loss where the cost of each class is inverse proportional to its prevalence. The score is the predicted probability of relevance:

$$s_t(x) = p(y_t = 1|x), \forall x \in S_2 \ . \tag{1}$$

3.2.2 Text-based Ensemble Ranker. We train an ensemble of text-based classifiers, each predicting the relevance of a reference with respect to a subset of R_1. The rationale is that a reference belongs to R_1 if it belongs to *any* subset of R_1. We construct subsets $\{R_g|g \in G, R_g \subset R_1\}$, each consisting of references cited in a group of semantically related sections of the 2013 ISA. For each group $g \in G$, we train a LR classifier on the training set $\{(x, y_g)|x \in S_1, y_g = \mathbf{1}\{x \in R_g\}\}$. The scoring function takes the highest predicted probability of all classifiers:

$$s_e(x) = \max_{g \in G} p(y_g = 1|x), \forall x \in S_2 \ . \tag{2}$$

3.2.3 Network-based Ranker. We represent each reference as nodes in a citation network. We retrieve citation relations from iCite, an NLM-maintained database [12]. Because iCite only covers PubMed articles, we take the subset of references $M \subset (S_1 \cup S_2)$ that have PMIDs. The network contains both PMIDs in M and those that cite or are cited by M. Node2vec [11] is used to learn 100-dimensional feature vectors for nodes (i.e., references in M). References without a PMID are assigned zero feature vectors. We train a LR classifier using $\{(x, y_n)|x \in S_1, y_n = \mathbf{1}\{x \in R_1\}\}$. The scoring function is:

$$s_n(x) = p(y_n = 1|x), \forall x \in S_2 \ . \tag{3}$$

3.2.4 Context Paragraph-based Ranker. This method learns a distance metric between a reference and a context paragraph. The idea is that if a reference is cited in a context paragraph, they should be close, and far apart otherwise. Formally, let P_1 be the set of context paragraphs in the 2013 ozone ISA. Let $Q^+ = \{(x, p)|x \in R_1, p \in P_1\}$ be the reference-paragraph pairs where x is cited in p. Let $Q^- = \{(x, p)|x \in S_1 - R_1, p \in P_1\}$ be a set of reference-paragraph pairs where x is not cited in p. Our goal is to construct a distance metric $d(\cdot, \cdot)$ such that $d(x, p)$ is small if $(x, p) \in Q^+$ and large if $(x, p) \in Q^-$. We first represent all references in S_1, S_2 and paragraphs in

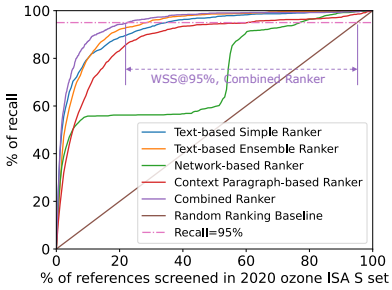

Figure 1: Recall curves of different ranking algorithms.

P_1 as 384-dimensional dense vectors using SBERT [15]. Then we learn a Mahalanobis distance metric $d_A(x,p) = \sqrt{(x-p)^\top A(x-p)}$ parameterized by a positive semi-definite matrix A. We find A by solving the following optimization problem [19]:

$$\min_{A \succeq 0} \sum_{(x,p)\in Q^+} d_A(x,p) \;\; s.t. \sum_{(x,p)\in Q^-} d_A(x,p) \geq 1 \; . \tag{4}$$

The score of a reference is its similarity to the nearest paragraph:

$$s_d(x) = \sigma\left(-\min_{p\in P_1} d_A(x,p)\right), \forall x \in S_2 \; . \tag{5}$$

$\sigma(\cdot)$ is the min-max normalization such that $0 \leq s_d(x) \leq 1$.

3.2.5 Combined Ranker. Aggregating ranking results from different methods often improves performance over individual methods [9]. We use a simple strategy to combine the previous four rankers by taking the average of their scores:

$$s_c(x) = (s_t(x) + s_e(x) + s_n(x) + s_d(x))/4, \forall x \in S_2 \; . \tag{6}$$

Figure 1 shows the recall curves of different rankers, including the random ranking baseline. The text-based ensemble ranker reaches 95% recall when 28.6% of S_2 are screened, the best among four individual rankers. The combined ranker achieves 95% recall when 22.4% of S_2 are screened, outperforming any individual ranker. The curve of network-based ranker has a bend because a large part of S_2 (56%) and R_2 (38%) do not have PMIDs and therefore cannot learn node vectors. It shows the limitation of constructing the citation network for these documents using only iCite.

3.3 Active Learning Experiments

We simulate the active learning process on S_2, the search result set of 2020 ozone ISA. In each iteration, we sample $k = 100$ references from S_2 and obtain their labels. This mimics the process of asking experts to review and judge the relevance of these references. We then update the same type of LR classifier as in Section 3.2.1 and rerank references in S_2 using predicted probability of relevance.

We evaluate ranking performance using work saved over sampling at recall R (WSS@R). It measures how much work (percentage of search results to be screened) is saved over random sampling to achieve a desired level of recall. The metric is commonly used in systematic review research [7, 14]. R is usually set to 95% to emphasize high recall. Given a ranked list, let TP (FP) be the number of true (false) positives above the cutoff rank that reaches 95% recall, and N be the total number of documents in the ranked list. Then $WSS@95\% = 95\% - (TP+FP)/N$. We illustrate WSS@95% for the combined ranker in Figure 1. It reaches 95% recall after screening 22.4% of S_2, saving (95-22.4)%=72.6% of work (or 72.6% $\times|S_2|$ =124,419 references) over random sampling. Clearly, the larger this metric, the better the ranked list.

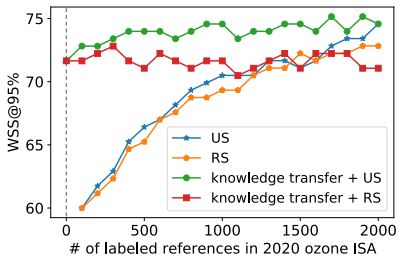

Figure 2: Learning curves of different active learning strategies. 'RS': random sampling; 'US': uncertainty sampling.

We use two methods to initialize the classifier. (1) Randomly sampling 100 references from S_2 to train an initial classifier. (2) Pseudo-labeling S_2 by **transferring knowledge** from the combined ranker $s_c(x)$. Let $r_c(x)$ be the rank position of reference $x \in S_2$ according to $s_c(x)$. We label the top-ranked $|R_1|$ references in S_2 as pseudo-relevant and train the initial classifier using $\{(x,y)|x \in S_2, y = \mathbf{1}\{r_c(x) \leq |R_1|\}\}$. In later iterations, the classifier is retrained on both pseudo labels and true labels. Loss on pseudo labels is downweighted by a factor of 10^{-2} to prevent it from dominating the loss function. Two sampling strategies are used in subsequent iterations. (1) **Random sampling**: selecting 100 references at random. (2) **Uncertainty sampling**: selecting 100 references predicted with probabilities closest to $p = 0.5$.

Figure 2 shows the learning curves of different active learning strategies. We observe a clear advantage of transferring the knowledge from a past ISA to the reviewing process of a later ISA on the same pollutant. The preliminary result highlights the missed opportunity by classical active learning methods: to drastically reduce screening efforts by exploiting the recurring nature of ISAs.

4 CONCLUSION AND FUTURE WORK

This paper introduces the literature screening dataset for the EPA's integrated science assessments (ISAs). The special nature of ISAs makes the literature screening task especially challenging: unprecedented scale, high-recall requirement, periodic updates, breadth of topics, and the policy relevance criterion. To support research on this problem, we provide three genres of data associated with two successive ISAs on the same pollutant. Our preliminary experiments show the promise of transferring relevance ranking knowledge learned from a previous ISA to a future ISA. The knowledge can be used to warm-start an active learning classifier, more effectively reducing the reviewing efforts in reaching a high recall than classical active learning classifiers that start from scratch.

Future work can not only explore more advanced machine learning techniques but also refine our problem formulations. For example, active learning strategies can be made methodical to provide high recall with statistical guarantees [8]. Various modalities of user feedback can be considered in active learning [4, 20], especially when the ultimate goal is *not* to train an accurate classifier, but to harvest all relevant documents [21]. Different years' ISAs may have topical shifts (e.g., the dosimetry chapter in the 2013 Ozone ISA was deprecated in the 2020 ISA), and therefore models learned from previous ISAs should account for distributional shift (e.g., covariate shift) when applied to later years [6].

REFERENCES

[1] United States Environmental Protection Agency. 2020. Integrated Science Assessment (ISA) for Ozone and Related Photochemical Oxidants (Final Report, April 2020).

[2] United States Environmental Protection Agency. 2022. Integrated Science Assessments (ISAs). https://www.epa.gov/isa. Accessed: 2022-04.

[3] Amal Alharbi and Mark Stevenson. 2019. A dataset of systematic review updates. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1257–1260.

[4] Josh Attenberg, Prem Melville, and Foster Provost. 2010. A unified approach to active dual supervision for labeling features and examples. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 40–55.

[5] Josh Attenberg and Foster Provost. 2011. Inactive learning? Difficulties employing active learning in practice. *ACM SIGKDD Explorations Newsletter* 12, 2 (2011), 36–41.

[6] Steffen Bickel, Michael Brückner, and Tobias Scheffer. 2009. Discriminative learning under covariate shift. *Journal of Machine Learning Research* 10, 9 (2009).

[7] Aaron M Cohen, William R Hersh, Kim Peterson, and Po-Yin Yen. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association* 13, 2 (2006), 206–219.

[8] Gordon V Cormack and Maura R Grossman. 2016. Scalability of continuous active learning for reliable high-recall text classification. In *Proceedings of the 25th ACM international on conference on information and knowledge management*. 1039–1048.

[9] Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. 2001. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*. 613–622.

[10] S Gopalakrishnan and P Ganeshkumar. 2013. Systematic reviews and meta-analysis: understanding the best evidence in primary healthcare. *Journal of family medicine and primary care* 2, 1 (2013), 9.

[11] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 855–864.

[12] B Ian Hutchins, Kirk L Baker, Matthew T Davis, Mario A Diwersy, Ehsanul Haque, Robert M Harriman, Travis A Hoppe, Stephen A Leicht, Payam Meyer, and George M Santangelo. 2019. The NIH Open Citation Collection: A public access, broad coverage resource. *PLoS biology* 17, 10 (2019), e3000385.

[13] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2019. CLEF 2019 technology assisted reviews in empirical medicine overview. In *CEUR workshop proceedings*, Vol. 2380.

[14] Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. 2016. Rayyan – a web and mobile app for systematic reviews. *Systematic reviews* 5, 1 (2016), 1–10.

[15] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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)*. 3982–3992.

[16] Harrisen Scells, Guido Zuccon, Bevan Koopman, Anthony Deacon, Leif Azzopardi, and Shlomo Geva. 2017. A test collection for evaluating retrieval of studies for inclusion in systematic reviews. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1237–1240.

[17] Burr Settles. 2009. Active learning literature survey. (2009).

[18] Hanna Suominen, Liadh Kelly, Lorraine Goeuriot, Aurélie Névéol, Lionel Ramadier, Aude Robert, Evangelos Kanoulas, Rene Spijker, Leif Azzopardi, Dan Li, et al. 2018. Overview of the CLEF eHealth evaluation lab 2018. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 286–301.

[19] Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. 2002. Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems* 15 (2002).

[20] Haotian Zhang, Gordon V Cormack, Maura R Grossman, and Mark D Smucker. 2020. Evaluating sentence-level relevance feedback for high-recall information retrieval. *Information Retrieval Journal* 23, 1 (2020), 1–26.

[21] Jie Zou and Evangelos Kanoulas. 2020. Towards question-based high-recall information retrieval: locating the last few relevant documents for technology-assisted reviews. *ACM Transactions on Information Systems (TOIS)* 38, 3 (2020), 1–35.