Noname manuscript No.
(will be inserted by the editor)

Automatic Document Screening of Medical Literature Using Word and Text Embeddings in an Active Learning Setting

Andres Carvallo · Denis Parra · Hans Lobel · Alvaro Soto

the date of receipt and acceptance should be inserted later

Abstract PRE-PRINT, to be published in Scientometrics `https://doi.org/10.1007/s11192-020-03648-6`.

Document screening is a fundamental task within Evidence-based Medicine (EBM), a practice that provides scientific evidence to support medical decisions. Several approaches have tried to reduce physicians' workload of screening and labeling vast amounts of documents to answer clinical questions. Previous works tried to semi-automate document screening, reporting promising results, but their evaluation was conducted on small datasets, which hinders generalization. Moreover, recent works in natural language processing have introduced neural language models, but none have compared their performance in EBM. In this paper, we evaluate the impact of several document representations such as TF-IDF along with neural language models (BioBERT, BERT, Word2Vec, and GloVe) on an active learning-based setting for document screening in EBM. Our goal is to reduce the number of documents that physicians need to label to answer clinical questions. We evaluate these methods using both a small challenging dataset (CLEF eHealth 2017) as well as a larger one but easier to rank (Epistemonikos). Our results indicate that word as well as textual neural embeddings always outperform the traditional TF-IDF representation. When comparing among neural and textual embeddings, in the CLEF eHealth dataset the models BERT and BioBERT yielded the best results. On the larger dataset, Epistemonikos, Word2Vec and BERT were the most competitive, showing that BERT was the most consistent model across different corpuses. In terms of active learning, an uncertainty sampling strategy combined with a logistic regression achieved the best performance overall, above other methods under evaluation, and in fewer iterations. Finally, we compared the results of evaluating our best models, trained using active learning, with other authors methods from CLEF eHealth, showing

DRAFT to be published in Scientometrics
https://doi.org/10.1007/s11192-020-03648-6

Pontificia Universidad Catolica de Chile

better results in terms of work saved for physicians in the document-screening task.

Keywords active learning · document screening · natural language processing

1 Introduction

Evidence-based Medicine (EBM) is a practice that provides scientific evidence to support medical decisions. This evidence nowadays is obtained from biomedical journals, usually accessible through the portal PubMed[1], a search engine which provides free access to abstracts of biomedical research articles, as well as to the MEDLINE database. An existing problem is to find relevant documents given a clinical question or a query within a massive volume of information. As a consequence, the time required for search and screening of articles can take long, and sometimes it consumes a large part of a physician's workday (Miwa et al., 2014; Elliott et al., 2014). When people conduct this repetitive task, there is a good chance of overlooking relevant articles, which can have a negative impact on decisions such as the patient's treatment (Keselman & Smith, 2012). Moreover, the publication of medical papers has grown exponentially in the last decade. Since 2005, PubMed has indexed more than 1 million articles per year, which means that the process of searching and manual screening of medical evidence will become increasingly more difficult for physicians without the support of information retrieval and machine learning algorithms. For this reason, some systems have emerged to support experts in the collection of evidence such as Embase[2], DARE[3] and Epistemonikos[4].

In this article, we present a method to improve the efficiency and efficacy of document screening in the practice of EBM. In other words, we aim at reducing the effort made by physicians at screening documents to find the evidence needed to support the answers of a medical question. Rather than building a classification model in the traditional machine learning way, where a large dataset of labeled documents is used to train a model, we choose to experiment with an active learning approach (Settles, 2012). We use active learning due to its similarity with the actual task carried upon by physicians in EBM: label a few documents in several iterations, and get better at classifying more documents after each iteration. One of the main tasks of active learning is choosing the appropriate data points (documents) to be labeled by the experts in order the train the model with as few examples as possible.

In order to evaluate our approach, we experiment with a large dataset of medical questions, unlike previous works that use smaller datasets (G. E. Lee & Sun, 2018). We aim to answer the following research question: Do a strategy based on state-of-the-art language models, such as BERT and BioBERT, in

[1] https://www.ncbi.nlm.nih.gov/pubmed/

[2] https://www.elsevier.com/solutions/embase-biomedical-research

[3] https://www.crd.york.ac.uk/CRDWeb/

[4] https://www.epistemonikos.org/en

DRAFT, to be published in Scientometrics
https://doi.org/10.1007/s11192-020-03648-6

conjunction with an active learning approach, helps to improve the efficiency and efficacy of document screening in the medical domain?. Furthermore, do these approaches represent a considerable advantage compared with traditional word embedding language models (Word2Vec and GloVe) and TF-IDF representation?

DRAFT, to be published in Scientometrics
https://doi.org/10.1007/s11192-020-03648-6

In this paper we contribute by:

1. experimenting in both a large dataset (Epistemonikos, 947 clinical questions) and a small dataset (CLEF eHealth, 50 clinical questions), showing evidence of generalization of our approaches,
2. comparing the performance of several document representations for active learning: TF-IDF and state-of-the-art language models' embeddings Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014)), BERT (Devlin et al., 2018) and BioBERT (J. Lee et al., 2019), a fine-tuned BERT for medical documents, as well as traditional relevance feedback,
3. validating the competitiveness of our method against other approaches used in the CLEF eHealth challenges over 2017 to 2019. Our method consistently tops other approaches in saving physicians' work for finding all relevant medical articles (total recall) given a medical question, and
4. sharing our code[5] and data[6] for research reproducibility.

2 Related Work

The task of finding relevant documents related to a medical question through citation screening has been studied and it is known as the *total recall problem*: given a medical topic or question, find all the documents that are relevant about a particular topic. Recently, the CLEF eHealth task 2 (Kanoulas et al., 2017, 2018, 2019) is a challenge that calls for solving the problem of prioritizing which documents to screen to reduce work overload for experts. They provide a public dataset with medical topics and a set of candidate documents; participants have to rank documents by relevance for every specific medical subject in the minimum of iterations to make more efficient the document screening process (Grossman et al., 2016).

In the literature, the approaches for solving this problem are based on three general lines: **information retrieval**, **machine learning methods**, and **natural language processing**. The latter is used to support the first two.

In the **information retrieval** area, there have been many attempts to solve the problem using techniques such as relevance feedback (Donoso-Guzmán & Parra, 2018), query expansion (G. E. Lee & Sun, 2018), ranking and inference based on external knowledge (Goodwin & Harabagiu, 2018).

From the **machine learning** community, the approaches usually are focused on semi-automate the screening process of medical articles, which is still conducted or validated by physicians. There have been efforts to solve this problem by using automatic classification (Bekhuis et al., 2014; Choi et al., 2012; Adeva et al., 2014; Mo et al., 2015; Wallace et al., 2012). In these previous works, authors compared classifiers such as Naive Bayes, K-NN, and SVM, using different ways to represent text, such as word embeddings and bag-of-clinical terms from titles and abstracts. There is also literature indicating the

[5] https://github.com/afcarvallo/active_learning_document_screening

[6] https://doi.org/10.5281/zenodo.3834845

DRAFT to be published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

use of active learning (Hashimoto et al., 2016; Figueroa et al., 2012; Wallace et al., 2010; Miwa et al., 2014) for medical topic detection and clinical text classification. Moreover, a few deep learning models have been proposed for the classification of relevant evidence and categorization of documents in medical questions (Del Fiol et al., 2018; Hughes et al., 2017). Generally, the majority of work done has used datasets of up to 50 medical topics/questions and 200,000 documents. In this work the dataset includes 948 medical questions and 370, 000 potential documents, allowing models to generalize and to improve their performance compared to the state of the art.

Moreover, for both machine learning and information retrieval approaches, there is an increasing use of more powerful Natural Language Processing techniques mainly derived from deep learning models (Peters et al., 2018; Devlin et al., 2018; Howard & Ruder, 2018).

The first generation of models for document representation were based on the vector space model (Salton et al., 1975) using TF-IDF vectors, but more recent approaches have represented words with models such as word and text embeddings. The methodology to obtain these embeddings has evolved, starting with Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and then full text embedding representations such as ELMO (Peters et al., 2018), ULM-fit (Howard & Ruder, 2018) and BERT (Devlin et al., 2018). The latter representation is state of the art in the field of language models, and it is based on the so called transformer architecture for neural networks (Vaswani et al., 2017) which includes attention mechanisms. BERT, for instance, predicts hidden words previously masked, and it also learns to predict sentences: if the second sentence in a pair of sentences is its subsequent in the original document or not. It can also be adapted to tasks such as text classification in the medical domain. For instance, Lee et al. (2019), re-trained BERT focusing on the biomedical domain with more than one million PubMed articles, thus generating a version of BERT called BioBERT. However, they did not test it for the task investigated in this article, document screening.

Within the last CLEF eHealth challenge for *Technology-Assisted Reviews*, participants used several approaches to address the problem of document screening: lexical statistics for relevant term identification (Alharbi & Stevenson, 2019), interactive BM25 (Di Nunzio, 2019), and a combination of ranking and a "greedy" sampling strategy to estimate the number of relevant documents (Li et al., 2019). In this paper, unlike previous work, we address the task with an active learning approach since it better reflects the work performed by the physicians while performing document screening. We take a different stand with respect to previous approaches, since our goal is to test the effectiveness of different document representations: bag-of-words (TF-IDF), neural word embeddings such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014)), as well as document language models such as BERT (Devlin et al., 2018) and BioBERT (J. Lee et al., 2019).

DRAFT, to be published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

3 Proposed Method

The process of finding documents that answer a clinical question requires first retrieving a set of candidate documents. Then, physicians perform the document screening where they select from the candidates abstracts and titles that are related to the medical question. This process may involve a large amount of time and cognitive effort from experts.
In this work, we propose the use of an active learning strategy to reduce the labeling effort from experts. Figure 1 illustrates the proposed approach.

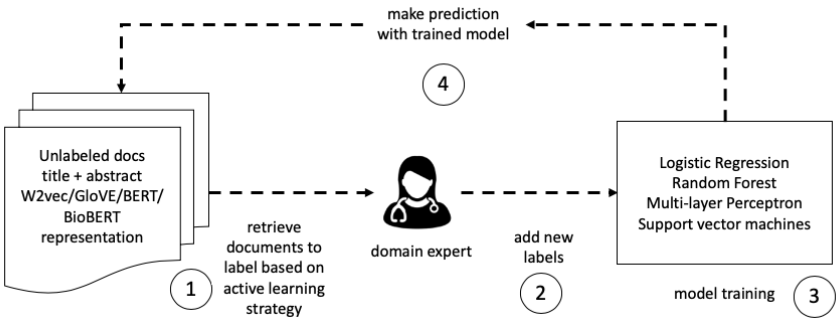


Fig. 1: Illustration of the active learning approach. It starts with a set of candidate documents which based on an active learning strategy (uncertainty or random sampling) are retrieved to be labeled. Then the oracle (domain expert) adds new labels and the system uses the labels to train a machine learning model, and next it makes predictions with the latest model trained. Predictions are used to sample the new set of candidate documents.

3.1 Efficient labeling using active learning

Given a medical question q, a set of unlabeled candidate documents $C = \{c_i\}_{i=1}^{N}$, and a labeling oracle O, in our case a physician who knows if a document is relevant to q, the goal of the process is to train a classifier of relevant documents M^q, using as few labelings from the oracle as possible. To achieve this, we iteratively select informative samples of documents to be labeled by the expert. Using these labelings, we progressively train the classifier, until we obtain a model with the desired performance, generating a sequence of classifiers $\{M_i\}_{i=1}^{k}$.
As the number of available oracle labelings is highly constrained, the critical aspect of the process is the selection of an appropriate sample to be presented to the oracle. To achieve this, we use an active learning approach (Settles, 2012), evaluating two different strategies for sampling, namely uncertainty sampling and random sampling. These strategies were selected based on their lower computational complexity compared to other methods such as error-based, gradient-based, and variable reduction (Settles, 2012). The first active learning strategy is uncertainty sampling, where one tries to select the sample

DRAFT, to be publishe
https://doi.org/10.1007/s1
metrics
3648-6

that the classification model is most uncertain about. Then to estimate this uncertainty, the scheme selects the sample with the lowest classification confidence when assigned to its most likely label. Formally, given an initial model Θ, we select a new sample $\hat{x}$ based on the following equation:

$$\hat{x} = \text{argmax}_x \, 1 - P_\theta(\hat{y}|x),$$

where $\hat{y}$ is the class label with the highest posterior probability given the classification model θ. The second strategy considered for experiments is random sampling. In this scheme, active learning randomly chooses examples to be labeled and then trains the model θ with these new labels.
Based on the selected sampling strategy, we obtain a small set of unlabeled documents $X = \{x_i, \}_{i=1}^{n}$ from C, with $n << N$. Following this, we query the oracle O for a binary labeling $Y = \{y_i, \}_{i=1}^{n}$ of the n examples in X, where $y_i = 1$ identifies relevant documents. Finally, using X and Y, we train a classification model $M_i(X, Y)$, that is used to predict the labels for unobserved documents. We repeat this process to create updated versions of the classification model M.

In practice, for the initial model M_1, we start with five randomly sampled labeled documents for each medical question and train the first version of the classification model.

3.2 Document representation

In this work, we compare TF-IDF representation (bag-of-words document representation with TF-IDF weighting) with word embeddings such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) , as well as with the state of the art text embedding BERT (Devlin et al., 2018), and a fine-tuned BERT model called BioBERT (J. Lee et al., 2019).
In order to train the word vectors, Word2Vec uses a feed-forward neural network for two possible tasks: given a sequence of words, predict the most probable next word (continuous bag of words) or given the word predict most probable context words (skip-gram). In this work, we use the Word2Vec skip-gram technique to obtain word embeddings, because it represents well even rare words (such as specific medical terms) compared to a continuous bag of words that presents higher accuracy for more frequent words (Mikolov et al., 2013).
In the case of GloVe, word embeddings are obtained based on a probabilistic approach. In this neural language model, the objective is that the dot product of a vector of a target word with a matrix of vectors from words of their context is as close as possible to the original word co-occurrence matrix. After that, when the vectors are already optimized using ordinary least squares, these word embeddings are used as a way to represent words in a latent space.
Concerning text embeddings such as BERT or BioBERT, they use a transformer architecture (Vaswani et al., 2017), an attention model that learns

DRAFT, to be published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

relations between words and sentences. As the transformer has an encoding and decoding architecture, in this case, BERT uses only the encoder. This language model reads all the sequence at once through a *query, key, value* structure and a positional encoding, using an attention mechanism to solve two tasks. The first task is predicting a hidden word, and the second aims to capture the relation between sentences, in this case, titles and abstracts of medical documents.

Our main goal in this article is to evaluate differences among models based on word embeddings and text embeddings. When using word embedding models (Word2Vec and GloVe) to represent a document, as shown in Figure 2, we have to aggregate the obtained embeddings from each word of the title and abstract to represent the document as a vector, and eventually use it as input in a machine learning model.

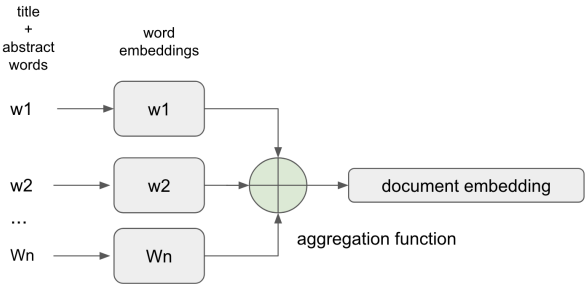


Fig. 2: Using Word embedding model (Word2Vec and GloVe) to transform the title and abstract words of an article into a single document embedding.

On the other hand, text embedding models such as BERT or BioBERT, as shown in Figure 3, take as input the complete document (title and abstract tokens) and independent of its length, they output a fixed-sized embedding that represents the document. Concerning BioBERT, it is a fine-tuned version of BERT with more than one million full-text documents from PubMed[7] and approximately 4.5 billion words. This model is adapted to the medical domain for tasks such as document classification.

To generate the document representations that serve as input for the active learning procedure, we employ the concatenation of title and abstract. As shown by G. E. Lee & Sun (2018), the combined information from title and abstract is more informative than each one of them separately. Once concatenated, we lowercase the text and remove stop-words. The resulting text is then processed by the selected embedding technique. For Word2Vec and Glove, a 300-dimension embedding vector is generated for each word, and the final representation is generated by averaging these vectors, ending up with a document vector of 300 dimensions. In the case of BERT and BioBERT, the whole text

[7] https://www.ncbi.nlm.nih.gov/pubmed/

DRAFT, to be published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

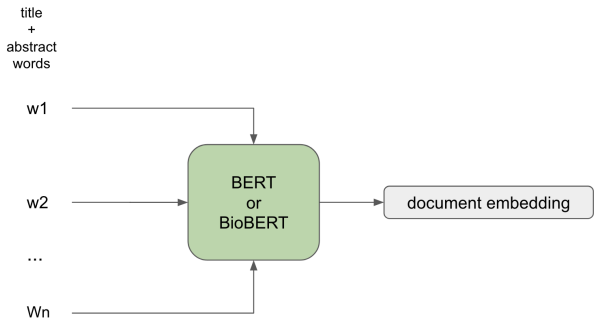


Fig. 3: Using text embedding model (BERT and BioBERT) to transform the title and abstract of an article into a document embedding.

is processed at once, generating a 768-dimension embedding vector as the final representation. Then for TF-IDF representation, we obtain a vector for each document and apply latent semantic indexing to the document-term matrix in order to reduce the dimensionality of each document to one hundred. If we do not perform this step, we might end up with document vectors of several thousands of dimensions (the size of the vocabulary), what might increase the chance of facing the *curse of dimensionality* when building the machine learning classification models. We have chosen the above embedding dimensions because it has been shown in several experiments that GloVE (Pennington et al., 2014) and Word2vec (Mikolov et al., 2013) achieve their best performance with embeddings of size 300. Moreover, for BERT-base (Devlin et al., 2018) and BioBERT (J. Lee et al., 2019), which was the one used in this case, the ideal dimension is 768, because we used a pre-trained BERT language model, which size is of 768 per document embeddings. Since we are using the optimal size for each language model (rather than the same dimension to all of them), we are giving them an equal chance of performance based on that parameter.

4 Datasets

To evaluate the proposed method, we use two datasets: CLEF eHealth[8] and Epistemonikos[9]. These datasets define a set of medical questions, where each is associated to a Systematic Review, which is a type of article that collects and synthesizes the relevant primary studies and trials related to a question. The information of each document in both datasets consists of the title, abstract, author, year and a label indicating if the document is relevant (or not) to the question or medical subject. For evaluation purposes, we split the documents related to each question (both relevant and not) into 70% for training and 30% for testing. We describe further characteristics of both datasets below.

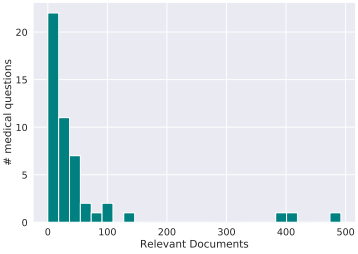
[8] `https://sites.google.com/site/clefehealth2017`

[9] https://www.epistemonikos.org/

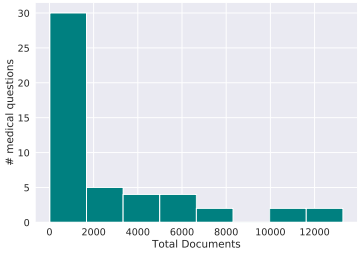
DRAFT, to be published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

Fig. 4: CLEF eHealth dataset distribution of relevant and total documents per question.

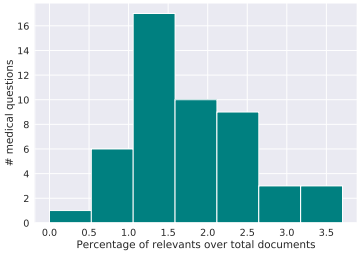
(a) Distribution of relevant documents per question.



(b) Distribution of total documents per question.



(c) Distribution of the percentage of relevant documents per question.



DRAFT, to be published in Scientometrics
https://doi.org/10.1007/s11192-020-03648-6

4.1 CLEF eHealth dataset

The CLEF eHealth dataset is conformed of 50 medical questions (ex. *which are the most effective treatments for the common cold?*) and 200,000 documents that were crawled from PubMed using each document id. Figure 4 presents the main characteristics of the distribution of the documents in the dataset: Figures 4(a) and 4(b) present the distribution of relevant documents and total documents per question in the CLEF eHealth dataset, respectively. On both, the y-axis represents the count of questions and the x-axis the number of documents. We can appreciate that most of the questions in CLEF eHealth have between 1 and 50 relevant documents, observing a long tail distribution. Regarding the total of documents, we can observe something similar since most of the questions have between 1 and 1000 documents. For instance, Figure 4(a) indicates that about 25 medical questions have between 1 to 10 relevant documents. The long bar indicates that this is the most frequent case. Then Figure 4(b) indicates that about 30 medical questions have between 1 and 1,800 documents (including relevant and not-relevant ones) which experts have to screen in order to identify relevant documents. So, plots a) and b) differentiate because a) shows the distribution of only relevant documents per

question and b) is the distribution of total documents including both relevant and not-relevant. Based on this, we argue that CLEF eHealth is a complex dataset because the proportion of relevant documents over the total number of documents is quite low. To assess this, Figure 4(c) presents the distribution of the proportion of relevant documents per question. It can be observed that most of the medical questions have a proportion between 1.5% and 2%, producing a highly unbalanced dataset. More details are shown in Table 8 in the Appendix.

4.2 Epistemonikos dataset

The Epistemonikos Evidence Synthesis Project is a collaborative initiative established in 2012 to collect, organize, and to compare all relevant evidence for health decision-making, through a multilingual platform. The resulting Epistemonikos dataset is composed of 948 medical questions and 372,829 potential documents. The labels were previously curated by senior medical students, in which they had to select papers related to a set of medical questions. Figure 5 presents the main characteristics of the distribution of the documents in the dataset:
Figures 5(a) and 5(b) present the distribution of relevant documents and total documents per question in the Epistemonikos dataset, respectively. In Figure 5(a) and 5(b) we have the distribution of relevant documents and the total documents in the Epistemonikos dataset. On both, the y-axis represents the count of questions and the x-axis the number of documents. We can appreciate that most of the questions in Epistemonikos have between 1 and 20 relevant documents, observing a long tail distribution. Regarding the total of documents, we can observe something similar since most of the questions have between 1 and 200 documents. For clarification, we provide an example. Figure 5(a) indicates that about 820 medical questions have between 1 and 20 relevant documents, then Figure 5(b) indicates that about 690 medical questions have between 1 and 200 documents (including the relevant ones) where experts have to screen to identify relevant documents. So, plots a) and b) differentiate because a) shows the distribution of only relevant documents per question and b) is the distribution of total documents including both relevant and not-relevant. Then, the proportion between relevant and total documents in this dataset is, on average, a 4.61%, which makes it less complex compared to CLEF eHealth. From Figure 5(c) we can be observe that most of the medical questions have a proportion between 4.8% and 5%. On appendix Table 9 we present more details of the proportion of relevant documents over the total in a sample of twenty medical questions.

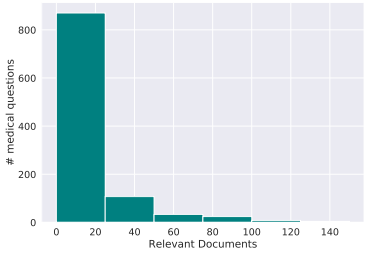
4.3 Epistemonikos and CLEF eHealth datasets complexity comparison

In this section, we compare the complexity of both datasets Epistemonikos and CLEF eHealth in terms of BM25 score similarity between medical questions

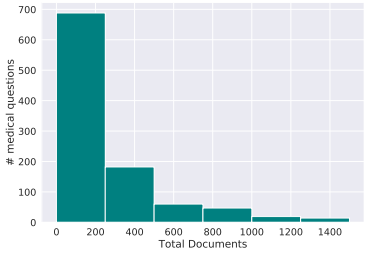
DRAFT to be Published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

Fig. 5: Epistemonikos test set distribution of relevant and total documents per question.

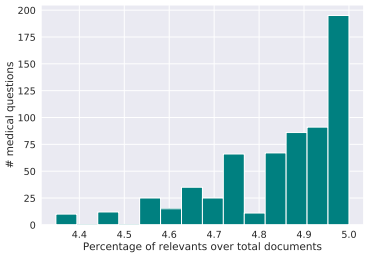
(a) Distribution of relevant documents per question.



(b) Distribution of total documents per question.



(c) Distribution of the percentage of relevant documents per question.



and their respective medical documents (Figure 6). Also, we calculate the proportion of medical terms over total words on each title and abstract from each dataset documents (Figure 7).

It can be seen from Figure 6 that most CLEF eHealth documents have a BM25 score between 0.1 and 0.2 compared to that of Epistemonikos, which is between 0.35 and 0.40. That indicates that the level of specificity and complexity of the CLEF dataset is higher for this task given by a lower chance to discriminate relevant documents only by the co-occurrence of words from the query and documents.

If we observe from Figure 7, the density of medical terms per document in CLEF eHealth, we see that it is higher than in Epistemonikos. Thus showing that the CLEF eHealth dataset has a vocabulary more focused on the medical domain, making it more complicated in the document screening task for the model learned since there is a larger probability of words unobserved during training to be used in testing data. CLEF eHealth texts have more medical terms compared to the Epistemonikos dataset. For BERT or BioBERT, it is easier than for GloVe or Word2vec to create meaningful aggregated document representations for the task addressed in this article.

DRAFT - to be published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

Fig. 6: Epistemonikos and CLEF eHealth comparison of BM25 query similarity

(a) Distribution of BM25 score on documents for CLEF eHealth queries

(b) Distribution of BM25 score on documents for Epistemonikos queries

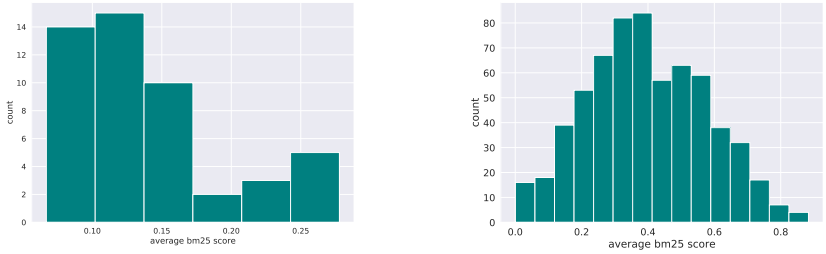
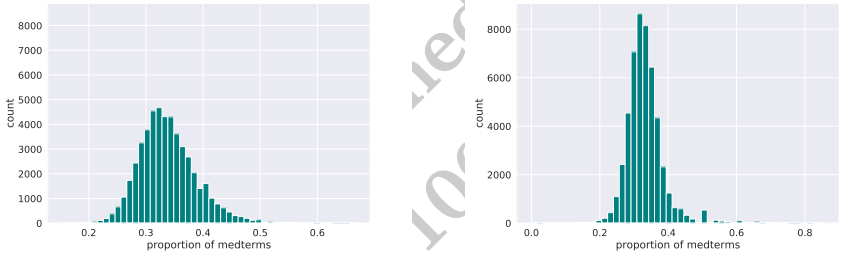


Fig. 7: Epistemonikos and CLEF eHealth comparison of medical terms proportion on document titles and abstracts

(a) Distribution of medical terms proportion on CLEF eHealth documents

(b) Distribution of medical terms proportion on Epistemonikos documents



5 Experimental evaluation

In this section, we compared the performance of combinations of different active learning strategies and documents representations. Experiments were programmed in Python3 using libact (Yang et al., 2017), sci-kit-learn (Pedregosa et al., 2011), pandas and gensim libraries. For tf-idf representation we used sci-kit-learn (Pedregosa et al., 2011) for feature extractor and reduced dimensionality with truncated SVD implementation.

In order to perform a large-scale evaluation, experiments are performed using a simulation of the active learning labeling process of documents for medical questions, using as the oracle the labels of the corresponding datasets.

Active learning setting: for each medical question, we hide the document labels and we leave only five random chosen documents with their respective labels to start building the model and then iterate with active learning to receive feedback from the oracle. For each prediction made by the machine

DRAFT, to be published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

learning model in each iteration, we sort the results depending on the predicted probability of being relevant for each model, so the evaluation metrics were calculated with the ranked list of potential candidates given by each strategy.
Relevance Feedback setting: we used two algorithms of relevance feedback Rocchio and BM25 as used by Donoso-Guzmán & Parra (2018) with the same meta parameters and setup but applied on this Epistemonikos dataset.
Classification models: we evaluate four different techniques for document relevance classification in our experiments: Multi-layer perceptron (MLP), Random Forest, Support Vector Machines (SVM), and Logistic Regression. These methods present a representative sample of machine learning techniques applied to text classification. The hyperparameters chosen for each method are, for Multilayer-Perceptron, three hidden layers of size 100, ReLU activation function, and Adam optimizer with a batch size of 32 (using a grid search optimization to obtain the best combination of hyperparameters). Then for Random Forest 100 estimators and a Gini Index criterion. Concerning Support Vector Machines, we used a radial basis function kernel and linear kernels with a regularization constant set at 1.0 (using a grid search optimization to tune it for the best hyperparameters). And finally, for Logistic Regression, we used an ℓ_2 regularization and a maximum of 100 iterations until convergence.
Evaluation metrics: we used traditional information retrieval metrics such as *recall@k*, *precision@k* and mean average precision (*MAP*), similar to G. E. Lee & Sun (2018). Also, non traditional metrics are used, such as *LastRel%* and *work saved over sampling (WSS)*, that were used as two-task submission evaluation metrics for CLEF eHealth 17 Competition (Goeuriot et al., 2017). *LastRel%* stands for last relevant percentage, which is the percentage of candidates documents that need to be screened and is essential because it indicates the number of documents needed to review to get all the relevant documents for that medical question. For example, if we have a list of 50,000 documents related to a medical question, where only 100 are relevant, the ideal would be that these 100 documents were in the top positions (last relevant in place 100) so that the expert did not have to review all 50,000 documents, indicating how efficient the proposed model is for solving this document screening task. Ideally, this metric should be as low as possible to avoid reviewing the entire list of articles until finding the last relevant document.
Justification of evaluation metrics: Recall@k indicates the ratio between between the retrieved relevant documents over the total relevant documents for a medical question. It is crucial because we do not want to miss any relevant document for a medical question. However, we need additional metrics because a naive optimization of recall and recall@k will make us find all the relevant documents (efficacy), but not in the most reasonable ranking (efficiency) to save physicians time. Then, precision@k calculates the proportion of relevant documents over k documents retrieved; it is still essential because we want to retrieve the maximum quantity of relevant documents on each iteration of the active learning loop. Mean Average Precision (MAP), computes precision at each recall positions (i.e., every position at which we find a relevant document) and averages over them. This metric penalizes a ranking that retrieves

DRAFT, to be published in Scientometrics
https://doi.org/10.1007/s11192-020-03648-6

relevant documents in positions too far away from the top. Finally, LastRel% reflects the number of medical documents that need to be revised until finding the latest relevant document for a specific medical question; this indicates if effectively the model is saving work from physicians. The same for WSS that reflects the amount of labor saved for the task of labeling relevant documents. Then *WSS* stands for work saved oversampling, which is a metric that shows how many candidate documents can be removed from manual screening. Evaluation metrics are related to the task of finding relevant documents for medical questions in the minimum possible iterations and on the first positions. Also, they allow evaluating if the proposed framework saves work to physicians for finding relevant documents without the need to review all available documents. Including both metrics (LastRel% and WSS) facilitates the comparison between models to verify the amount of work saved from physicians thanks to the proposed framework, for the task of document screening related documents to medical questions.

6 Results

6.1 CLEF eHealth dataset results

For these experiments we evaluated the active learning framework combining document representation, active learning strategies and machine learning models for a small dataset (CLEF eHealth). The results shown are recall at three cut off levels (recall@10, recall@20, recall@30), precision at three cut off levels (10,20,30), mean average precision, Lastrel% and WSS after ten labeling iterations of ten documents each. In *Table 1*, we see the results on a small dataset (CLEF eHealth).

In Table 1, the first column indicates the dataset as well as the type of embeddings. The second column shows the active learning strategy (US vs. RS), as well as the learning model (MLP, RF, LR, SVM). Then the following nine columns show recall at three cut off levels (recall@10, recall@20, recall@30), precision at three cut off levels (precision@10, precision@20, precision@30), Mean average precision (MAP), Lastrel% and WSS. The larger these metrics, the better the model, except for Lastrel% (the smaller, the better). As shown in Table 1 for the CLEF eHealth dataset, the combination of random forest (RF) with an uncertainty sampling (US) strategy and BioBERT representation achieves the best performance in recall@k, and the best in precision@k. However, there are no significant differences with the results obtained using BERT with RF and BioBERT with SVM-linear. When comparing state-of-art representations (BERT, BioBERT) with word embeddings and TF-IDF representations, we noticed that although these representations do not report the best results, they are more consistent and robust to changes in the sampling strategy.

If we look at the latest relevant documents rather than the top-k, we see an interesting result. Concerning work saved over the document screening task,

DRAFT, to be published in Scientometrics
https://doi.org/10.1007/s11192-020-03648-6

Table 1: Average results of recall@k (r@k), precision@k (pr@k), Mean Average Precision (MAP), Lastrel% and WSS performance measured in CLEF eHealth dataset using active learning strategies (US: uncertainty sampling, RS: random sampling) using a batch of 10 documents per feedback iteration for TF-IDF, Word2vec, GloVe, BERT-base and BioBERT-base representation. Results in **bold** font are the best for each metric, while the second and third best are underlined. Statistical significance is calculated with multiple t-tests using Bonferroni correction. The symbol $*$ indicates the statistically significant best result for each metric, but results with no significant difference with the best one are indicate with †.

Dataset	AL-Model	r@10	r@20	r@30	pr@10	pr@20	pr@30	MAP	LastRel%	WSS
CLEF eHealth	US-MLP	.081	.120	.163	.173	.151	.133	.128	85.9	.141
50 SRs	US-RF	.334	.392	.418	.367	.408	.320	.404	76.9	.231
TF-IDF	US-LR	.255	.320	.355	.414	.308	.241	.278	75.7	.243
	US-SVM (rbf)	.292	.335	.364	.471	.313	.241	.327	74.7	.206
	US-SVM (linear)	.268	.310	.331	.453	.313	.246	.315	77.6	.224
	RS-MLP	.034	.067	.097	.072	.076	.076	0.07	85.4	.145
	RS-RF	.167	.227	.295	.293	.221	.196	.211	76.1	.238
	RS-LR	.126	.189	.242	.238	.201	.187	.175	75.0	.249
	RS-SVM (rbf)	.116	.180	.224	.255	.207	.184	.178	78.3	.216
	RS-SVM (linear)	.144	.212	.280	.255	.221	.197	.205	73.9	.260
CLEF eHealth	US-MLP	.132	.200	.221	.233	.173	.133	.176	76.0	.240
50 SRs	US-RF	.223	.281	.313	.341	.219	.165	.266	83.0	.170
GloVe	US-LR	.263	.311	.330	.396	.256	.188	.290	64.8	.352
300 dim	US-SVM (rbf)	.228	.265	277	.341	.213	.148	.247	74.0	.260
	US-SVM (linear)	.239	.274	.289	.343	.221	.160	.260	76.0	.240
	RS-MLP	.122	.139	.198	.225	.155	.122	.154	78.8	.212
	RS-RF	.128	.183	.218	.203	.150	.122	.156	84.0	.160
	RS-LR	.113	.186	.247	.185	.156	.143	.161	74.9	.251
	RS-SVM (rbf)	.144	.227	.279	.226	.174	.142	.181	96.1	.039
	RS-SVM (linear)	.126	.181	.240	.187	.145	.119	.146	68.1	.318
CLEF eHealth	US-MLP	.215	.264	.278	.322	.220	.167	.252	66.8	.332
50 SRs	US-RF	.265	.308	.330	.382	.245	.184	.297	74.9	.251
Word2vec	US-LR	.228	.266	.278	.345	.215	.160	.252	68.0	.320
300 dim	US-SVM (rbf)	.237	.286	.308	.412	.278	.205	.293	71.6	.284
	US-SVM (linear)	.235	.272	.279	.394	.249	.177	.259	73.2	.268
	RS-MLP	.118	.173	.232	.179	.156	.134	.166	77.1	.229
	RS-RF	.144	.187	.256	.197	.137	.119	.170	83.0	.170
	RS-LR	.102	.167	.252	.137	.121	.121	.118	74.0	.260
	RS-SVM (rbf)	.121	.180	.238	.191	.156	.133	.160	85.3	.147
	RS-SVM (linear)	.164	.226	.249	.173	.145	.118	.162	69.0	.329
CLEF eHealth	US-MLP	.481	.663	.762	.802	.688	.597	.816	12.9	.871
50 SRs	US-RF	.565†	.727†	.804†	.833†	.695†	.597†	.893†	6.2	.938
BERT-base	US-LR	.561†	.721†	.800†	.837†	.693†	.591†	.852†	9.8	.902
768 dim	US-SVM (rbf)	.560†	.705	.783	.835†	.678	.579	.826	22.1	.779
	US-SVM (linear)	.570†	.736†	.813†	.841†	.706†	.601†	.876	13.4	.866
	RS-MLP	.082	.125	.174	.106	.099	.089	.108	80.6	.194
	RS-RF	.130	.165	.189	.141	.107	.080	.130	83.9	.161
	RS-LR	.178	.271	.320	.272	.219	.181	.214	73.1	.269
	RS-SVM (rbf)	.165	.232	.288	.194	.158	.137	.173	89.7	.103
	RS-SVM (linear)	.147	.212	.248	.183	.143	.128	.168	67.6	.323
CLEF eHealth	US-MLP	.486	.667	.758	.806	.697	.604	.840	12.0	.880
50 SRs	US-RF	**.571***	**.738***	**.819***	**.853***	**.715***	**.614***	**.910***	**4.5***	**.955***
BioBERT-base	US-LR	.559	.723	.805	.831	.696	.595	.855	9.5	.905
768 dim	US-SVM (rbf)	.555	.702	.781	.824	.677	.577	.822	18.9	.811
	US-SVM (linear)	.571†	.736†	.815†	.841†	.706	.603	.881†	12.2	.878
	RS-MLP	.126	.174	.225	.139	.113	.105	.140	81.1	.189
	RS-RF	.111	.142	.177	.191	.133	.114	.142	86.7	.133
	RS-LR	.201	.254	.290	.219	.165	.138	.216	70.5	.295
	RS-SVM (rbf)	.187	.248	.280	.232	.180	.146	.205	86.4	.136
	RS-SVM (linear)	.176	.243	.273	.216	.174	.140	.203	67.8	.321

DRAFT, to be published in Scientometrics
https://doi.org/10.1007/s11192-020-03648-6

the RF model combined with a BioBERT representation, with an uncertainty sampling strategy, has the best performance, since the expert would have to review on average only a five percent (4.5%) of the list until finding the last

relevant document. In contrast, with GloVe representation using random sampling, but with an SVM with RBF kernel as the learning method, the expert had to review an average of 96.1% of the full list.

6.1.1 CLEF eHealth model learning analysis

In this section, we present the results for the CLEF eHealth dataset of recall@10 after each iteration of documents for machine learning models (Multilayer-Perceptron, Random Forest, Support Vector Machines with linear kernel and Logistic Regression). We considered the best three document representations results obtained for the CLEF eHealth dataset (BioBERT, BERT, and GloVe). For each representation (*Figures 8-10*), we have a comparison of uncertainty-based active learning with random sampling. On the x-axis, we have the number of iterations of ten documents that we ask the oracle to label, and on the y-axis is the metric of recall@10 on each iteration. For this particular task of document screening, on the iteration analysis, we focused on the recall@10, because our goal is not to leave out relevant documents for a medical question. Figures 8, 9 and 10 show that BioBERT document representation for CLEF eHealth dataset gets higher levels of effectiveness at tenth iteration. Also, with BioBERT document representation, Logistic Regression and Random Forests gets better results in fewer iterations and are a clear winners over other models. Regardless of how we represent the documents and the machine learning model, the strategy of active learning based on uncertainty surpasses the baseline random sampling in all cases.

6.2 Epistemonikos dataset results

For these experiments, we evaluated our active learning framework on a large dataset of questions (Epistemonikos, 948 questions) combining document representations, active learning strategies, and machine learning models. Similar to CLEF eHealth, we used the same evaluation metrics to make them comparable. We also compared our results with traditional relevance feedback algorithms (using BM25 and Rocchio), using the same setting as Donoso-Guzmán & Parra (2018) but applied on this dataset.

Table 2 presents the results for the Epistemonikos dataset. The first column indicates the dataset as well as the type of embeddings. The second column shows the active learning strategy, as well as the learning model. Later, the following nine columns then show recall at three cut off levels (recall@10, recall@20, recall@30), precision at three cut off levels (precision@10, precision@20, precision@30), Mean average precision (MAP), Lastrel% and WSS.

As shown in Table 2 for the Epistemonikos dataset, it can be seen that the combination of an uncertainty sampling strategy with a logistic regression (LR) using a Word2vec representation of documents achieves the best results in terms of performance at recall@10. However, there is not a major improvement over GloVe representation using the same model and active learning

DRAFT, to be published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

Fig. 8: Comparison of Uncertainty and Random Sampling performance for CLEF eHealth dataset using BioBERT document representation, iterations versus recall@10.

(a) Uncertainty Sampling **(b)** Random Sampling

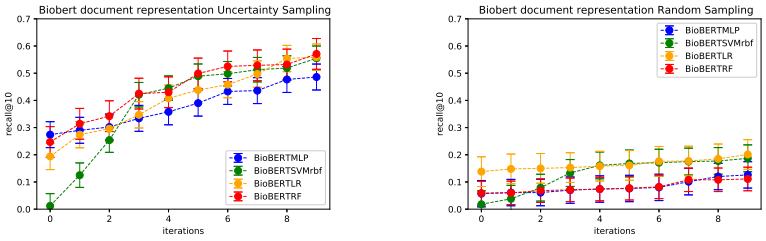


Fig. 9: Comparison of Uncertainty and Random Sampling performance for CLEF eHealth dataset using BERT document representation, iterations versus recall@10.

(a) Uncertainty Sampling **(b)** Random Sampling

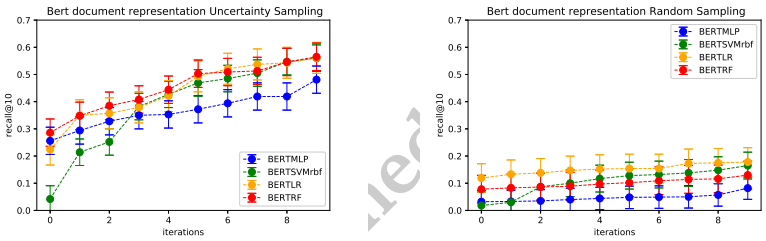
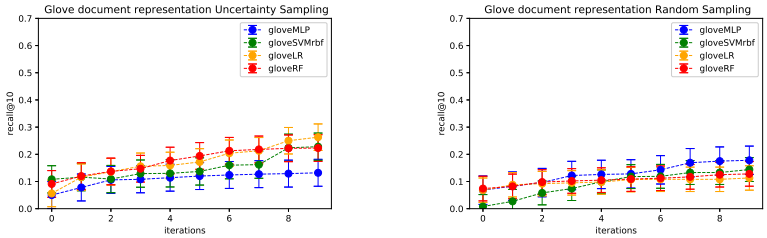


Fig. 10: Comparison of Uncertainty and Random Sampling performance for CLEF eHealth dataset using GloVe document representation, iterations versus recall@10.

(a) Uncertainty Sampling **(b)** Random Sampling



strategy, and there are no significant differences compared to SVM and MLP. Concerning work saved oversampling (WSS), the LR model combined with a Word2vec representation has the best performance since the expert would have to review, on average, only 14.8% of the list until finding the last relevant document.

With this same Word2vec representation, we see an excellent performance of US-MLP in terms of recall@k and precision@k metrics, indicating that MLP

Table 2: Average results of recall@k (r@k), precision@k (pr@k), Mean Average Precision (MAP), Lastrel% and WSS performance measured in Epistemonikos dataset using active learning strategies (US: uncertainty sampling, RS: random sampling), with batch size of 10 documents per feedback iteration for TF-IDF, Word2vec, GloVe, BERT-base and BioBERT-base representation. Results in **bold** font are the best for each metric, while the second and third best are underlined. Statistical significance by multiple t-tests using Bonferroni correction. The symbol $*$ indicates the statistically significant best result for each metric, but results with no significant difference with the best one are indicated with †.

Dataset	AL-Model	r@10	r@20	r@30	pr@10	pr@20	pr@30	MAP	LastRel%	WSS
Epistemonikos	US-MLP	.242	.347	.434	.294	.215	.173	.255	75.5	.245
948 SRs	US-RF	.516	.587	.630	.534	.347	.262	.518	68.4	.290
TF-IDF	US-LR	.507	.591	.633	.517	.333	.247	.477	66.7	.333
	US-SVM (rbf)	.441	.513	.552	.442	.281	.207	.416	68.7	.313
	US-SVM (linear)	.483	.556	.600	.491	.317	.235	.460	67.7	.323
	RS-MLP	.143	.227	.313	.110	.091	.083	.130	76.5	.234
	RS-RF	.380	.468	.527	.366	.246	.189	.345	70.5	.294
	RS-LR	.428	.531	.589	.399	.279	.218	.392	63.2	.367
	RS-SVM (rbf)	.428	.515	.569	.392	.265	.205	.391	64.1	.358
	RS-SVM (linear)	.433	.525	.582	.406	.278	.213	.402	64.2	.357
Epistemonikos	US-MLP	.508	.666	.744	.555	.421	.337	.591	32.2	.678
948 SRs	US-RF	.694	.832	.884	.696	.497	.385†	.765	23.5	.765
GloVe	US-LR	.706†	.844†	.898†	.689	.494	.385†	.768	15.3	.847
300 dim	US-SVM (rbf)	.697	.828	.877	.670	.470	.361	.744	17.9	.821
	US-SVM (linear)	.704†	.841†	.896	.693	.495	.384†	.772	16.1	.839
	RS-MLP	.538	.673	.737	.439	.319	.253	.492	60.2	.398
	RS-RF	.573	.694	.754	.483	.338	.265	.522	49.1	.509
	RS-LR	.684	.814	.866	.589	.419	.329	.668	33.3	.667
	RS-SVM (rbf)	.707	.830	.877	.616	.436	.340	.705	72.0	.280
	RS-SVM (linear)	.708	.832	.877	.619	.437	.338	.708	31.0	.690
Epistemonikos	US-MLP	.714†	**.854***	**.903***	**.707**	**.504***	**.392***	**.787***	16.1	.839
948 SRs	US-RF	.695	.832	.888	.703†	.501	.388†	.765	23.5	.765
Word2vec	US-LR	**.717***	.851†	.900†	.697	.492	.381	.768	**14.8***	**.852***
300 dim	US-SVM (rbf)	.705†	.844†	.898†	.698	.496	.385†	.769	16.3	.837
	US-SVM (linear)	.706†	.835	.889	.688	.489	.379	.763	18.1	.819
	RS-MLP	.694	.821	.872	.605	.431	.338	.692	33.0	.670
	RS-RF	.568	.699	.764	.487	.348	.275	.525	48.6	.514
	RS-LR	.676	.807	.864	.579	.416	.329	.658	35.0	.650
	RS-SVM (rbf)	.705	.832	.878	.619	.439	.342	.709	70.7	.293
	RS-SVM (linear)	.694	.817	.868	.604	.428	.334	.691	33.3	.667
Epistemonikos	US-MLP	.514	.685	.771	.577	.428	.335	.577	37.7	.623
948 SRs	US-RF	.669	.802	.856	.673	.473	.364	.718	32.5	.675
BERT-base	US-LR	.705†	.834	.883	.702†	.494	.381	.767	21.9	.781
768 dim	US-SVM (rbf)	.685	.814	.864	.680	.476	.361	.733	26.0	.74
	US-SVM (linear)	.692	.825	.876	.701†	.496	.380	.755	24.7	.753
	RS-MLP	.411	.542	.621	.326	.239	.194	.342	68.1	.319
	RS-RF	.486	.614	.684	.393	.282	.225	.410	62.4	.376
	RS-LR	.645	.767	.824	.566	.396	.310	.623	47.9	.521
	RS-SVM (rbf)	.652	.764	.821	.567	.393	.306	.626	73.1	.269
	RS-SVM (linear)	.647	.765	.815	.559	.389	.303	.628	29.9	.700
Epistemonikos	US-MLP	.450	.612	.695	.518	.389	.309	.513	42.5	.575
948 SRs	US-RF	.443	.587	.674	.422	.307	.246	.411	50.2	.498
BioBERT-base	US-LR	.673	.806	.868	.656	.463	.359	.712	23.2	.768
768 dim	US-SVM (rbf)	.664	.797	.853	.651	.456	.353	.695	26.5	.735
	US-SVM (linear)	.666	.794	.850	.641	.447	.343	.691	26.3	.737
	RS-MLP	.469	.603	.676	.393	.285	.228	.418	72.9	.271
	RS-RF	.557	.684	.750	.470	.335	.264	.503	57.4	.426
	RS-LR	.690	.812	.860	.604	.427	.333	.681	38.4	.616
	RS-SVM (rbf)	.681	.804	.852	.597	.422	.329	.674	76.8	.232
	RS-SVM (linear)	.683	.803	.848	.596	.418	.323	.671	22.6	.773
Epistemonikos	Rocchio	.261	.369	.432	.655	.419	.330	.631	26.31	.737
948 SRs	BM25	.131	.173	.209	.427	.295	.240	.254	67.24	.328

DRAFT, to be published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

performs well at ranking the top documents. Concerning a general comparison between active learning versus relevance feedback approaches (Rocchio and BM25, at the end of Table 2), regardless of the representation of documents,

Fig. 11: Comparison of Uncertainty and Random Sampling performance for Epistemonikos dataset using Word2vec document representation, iterations versus recall@10.

(a) Uncertainty Sampling **(b)** Random Sampling

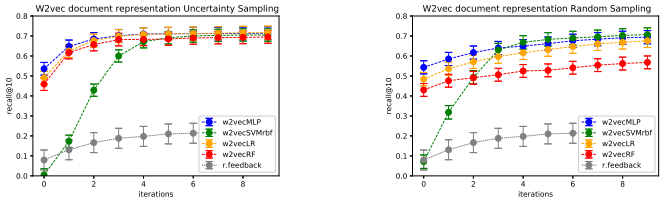


Fig. 12: Comparison of Uncertainty and Random Sampling performance for Epistemonikos dataset using GloVe document representation, iterations versus recall@10.

(a) Uncertainty Sampling **(b)** Random Sampling

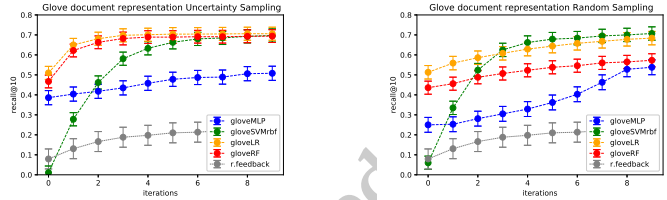
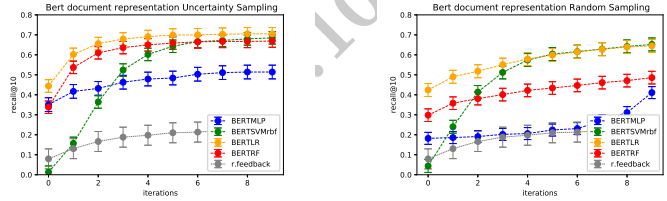


Fig. 13: Comparison of Uncertainty and Random Sampling performance for Epistemonikos dataset using BERT document representation, iterations versus recall@10.

(a) Uncertainty Sampling **(b)** Random Sampling



machine learning models, or active learning strategy, a definite improvement can be observed on all metrics.

6.2.1 Epistemonikos model learning analysis

In this section, we present the results for the Epistemonikos dataset of recall@10 after each iteration of the active learning process with ten documents per iteration. The machine learning models (Multilayer-Perceptron, Random Forest, Support Vector Machines with linear kernel and Logistic Regression) are compared on the best three document representations results obtained for this dataset (Word2Vec, GloVe, and BERT), all of them compared to the base-

DRAFT, https://doi.org/10.1007/s11192-020-03648-6

line relevance feedback model (Rocchio). For each representation, we have a comparison of uncertainty based active learning with random sampling. On the x-axis, we have the number of iterations of ten documents that we ask the oracle to label, and on the y-axis is the metric of recall@10 on each iteration. Figures 11 - 13 show that for the Epistemonikos dataset, all methods converge more quickly with uncertainty sampling than with random sampling, which saves a considerable deal of effort to physicians for labeling. Moreover, Word2Vec embedding representation seems to speed up convergence compared to GloVe and BERT on several learning methods (notice the effect on SVM). However, there are no significant differences in Word2Vec with GloVe or BERT after ten iterations (also shown in Table 2). In all cases, the Logistic Regression (LR) and Random Forest (RF) models reports higher levels of recall than the other methods from iteration one and converges after the 3rd or 4th iteration, which is in deep contrast to SVM or MLP which converge only at the 7th or 4th iteration. This result provides essential evidence of the effort that could be saved to physicians as oracles, with only 40 documents labeled rather than 60 or 70 to achieve a similar level of classifier performance by using uncertainty sampling with logistic regression for the sampling strategy and learning algorithm, respectively. Finally, using an uncertainty-based sampling strategy, independent of model or representation of the document that we use, we outperform the relevance feedback baseline very quickly compared to random sampling.

7 Comparison with participants from CLEF eHealth challenge

In order to better position our approach against competing methods, we compared it against the runs of a formal CLEF competition. In this section, we contrast our best models, representation, and active learning strategy combinations with the other participants of CLEF eHealth challenges over 2017-2019 task 2 (Technologically Assisted Reviews in Empirical Medicine[10]), which consists of screening MEDLINE abstracts to identify relevant articles to a medical issue within a set of candidates.

We evaluated all the submissions of the CLEF eHealth participants and picked their best result to compare them fairly against ours, using the official evaluation scripts provided in the challenge web repository[11].

For this comparison, we extended our original model to be able to use the CLEF eHealth evaluation scripts and generalize to new medical questions non-observed in the training set, without requiring additional training data. Now, given a medical question q, the input of the model are tuples made of the question concatenated with the document to classify d, i.e., $M(q, d)$. In this way, we train only one global model M instead of one model for each question M^q, as described in section 3.1. The number of labels requested to the oracle O stays the same as in the original model. In summary, we do not alter the

[10] https://sites.google.com/site/clefehealth2017/task-2

[11] https://github.com/CLEF-TAR/tar

DRAFT to be published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

document representation, but rather the input of the model from $M(d)$ to $M(q, d)$.

For obtaining each query-document representation, we apply BERT-base and BioBERT-base to the concatenation of both texts as the input sequence. We use padding to get a maximum sequence length of 512 characters, and the output embedding representing our query-document corresponds to the "[CLS]" token embedding in the last layer (768 dimensions). This setting has yielded good results in text ranking for general-domain documents (Qiao et al., 2019; Nogueira et al., 2019), but it has not been tested for biomedical literature. Besides, we add information about the relationship between question and document pairs (q, d) using BERT separator indicators.

We also introduce a step into the active learning process, suggested by Nogueira et al. (2019). Before doing the first iteration of document prediction to perform uncertainty sampling, we select a percentage of the potentially most relevant documents using BM25. After some experiments, we concluded that 65% of the documents were the most appropriate percentage. Then, we apply uncertainty sampling over this pre-filtered set and continue the traditional active learning process.

In order to report the results, we use the same evaluation metrics as CLEF eHealth, described as follows:

- **Lastrel:** this metric shows the index of the last relevant document found, indicating the minimum number of documents returned to retrieve all relevant documents.
- **Work saved over sampling (wss100 and wss95):** the amount of work saved by physicians for the document screening task.
- **Average precision: (ap)** a combination of precision and recall for ranked retrieved documents.
- **Normalized area under the precision-recall curve: (norm_area)** this metric shows area under the cumulative recall curve normalized by the optimal area.
- **Normalized cumulative gain at k% (ncg@20, ncg@40, ncg@60):** this shows recall at a k percentage of shown documents.

7.1 Benchmark CLEF eHealth 2017

In this section we compare the performance of our model with other participants from CLEF eHealth 2017. Below we briefly describe the methods used by the other participants from CLEF eHealth 2017 proceedings (Kanoulas et al., 2017):

- **IIT:** used a query expansion approach based on relevance feedback and on TF-IDF similarity (J. Singh & Thomas, 2017).
- **ECNU:** proposed a learning to rank approach that combines BM25, PL2, BB2 as features; and then for the trained model they included a vector space model (Chen et al., 2017).

DRAFT, to be published in Scientometrics
https://doi.org/10.1007/s11192-020-03648-6

- **ETH:** used a LAMBDA-Mart model trained on features, such as BM25, Fuzzy search, and vector content representation (Hollmann & Eickhoff, 2017).
- **NCSU:** they adopted an active learning approach using an SVM classifier. Representing documents with TF-IDF (Yu & Menzies, 2017).
- **Waterloo:** applied improvements to the Baseline Model Implementation (BMI) from the TREC total recall track 2015-2016. After that, they used the "knee-method" stopping criteria (Cormack & Grossman, 2016) to BMI to determine which documents have to be revised by the expert (Cormack & Grossman, 2017), with the purpose of achieving high recall and high probability.
- **QUT:** trained a learning to rank model using PICO (population, intervention, control, outcome) questions features (Scells et al., 2017).
- **UOS:** compared two models: Latent Dirichlet Allocation (LDA) and Roccio relevance feedback (Kalphov et al., 2017).
- **AUTH:** used a learning to rank approach combining batch and active learning (Lagopoulos et al., 2018).
- **CNRS:** trained a logistic regression from n-gram features using the title, abstract, and Medline citations (C. Norman et al., 2017).
- **Padua:** used a two-dimensional probabilistic version of BM25 to rank documents (Di Nunzio et al., 2017).
- **Sheffield:** parsed the boolean queries to extract terms and MesH headings and used TF-IDF cosine similarity to calculate the similarity between medical questions and documents (Alharbi & Stevenson, 2017).
- **AMC:** trained a random forest over topic model representation of the documents (van Altena & Olabarriaga, 2017).
- **NTU and UCL:** both trained a deep neural network to identify articles relevant to medical questions. For more details on the architecture see G. Singh et al. (2017); G. E. Lee (2017)

The results in Table 3 indicate that in terms of last relevant document (lastrel) our method US-RF-BioBERT is the best among all. For a task involving total recall, this result implies that we are better than any other method in reducing the workload of physicians to review a large portion of documents to find all the scientific evidence.

Wss100 and wss95 metrics give more evidence of this result. In terms of work saved over sampling wss100, our method US-RF-BioBERT is ranked second and US-RF-BERT third, respectively. These methods also perform very competitively in terms of wss95, reaching third and fourth positions, respectively. Although *BMI combined with knee method* (Waterloo) shows better than us in terms of wss, but with the cost of more than double the value of lastrel (1464) than our best approach US-RF-BioBERT (531).

Finally, as for normalized cumulative gain ($ncg@k$), with our best model US-RF-BioBERT, we reach only a seventh place in terms of ncg@20 (.712), but we improve to 4th position in ncg@40 (.892) and ncg@60 (.963), and we reach the third place in terms of norm_area. As expected, average precision is a metric

DRAFT, to be published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

Table 3: Benchmark of our best active learning strategies, models and document representations with other participants and baselines from CLEF eHealth 2017

model	lastrel	wss100	wss95	ncg@20	ncg@40	ncg@60	norm_area	ap
US-RF BioBERT	**531**	.658	.633	.712	.892	.963	.862	.211
IIT	548	.111	.135	.390	.454	.462	.675	.159
ECNU	725	**.770**	.175	.425	.471	.494	.651	.166
US-RF BERT	740	.638	.597	.556	.764	.868	.812	.142
ETH	785	.122	.163	.666	.737	.756	.744	.207
NCSU	928	.145	.268	.765	.844	.861	.684	.110
US-LR BioBERT	1061	.506	.472	.491	.723	.816	.747	.085
US-LR BERT	1458	.411	.396	.449	.660	.782	.715	.082
Waterloo	1464	.611	**.701**	**.887**	**.974**	**.995**	**.927**	.318
QUT	1873	.110	.129	.478	.647	.709	.619	.120
UOS	1857	.221	.348	.547	.766	.861	.727	.124
AUTH	2119	.519	.690	.868	.962	.985	.920	.293
CNRS	2250	.412	.497	.717	.887	.955	.839	.179
Padua	2260	.398	.496	.816	.913	.945	.885	.269
Sheffield	2382	.395	.488	.691	.889	.967	.847	.218
Baseline BM25	2851	.285	.400	.645	.828	.927	.809	.174
AMC	2913	.249	.333	.386	.790	.899	.761	.129
NTU	3570	.091	.075	.173	.394	.586	.538	**.520**
Baseline Rndm	3722	.040	.034	.192	.379	.575	.484	.045
UCL	3801	.072	.064	.229	.440	.627	.507	.060

where we do not perform very well (5th place), but we consider this a trade-off of having the best performance in terms of lastrel for a task focused on total recall.

7.2 **Benchmark CLEF eHealth 2018**

In this section we compare the performance of our model with other participants from CLEF eHealth 2018. Below we briefly describe the methods used by the other participants from CLEF eHealth 2018 proceedings (Kanoulas et al., 2018):

- **Waterloo:** they used the same method proposed in CLEF eHealth 2017 (Cormack & Grossman, 2017).
- **UNIPD:** proposed a two dimensional BM25 approach (Di Nunzio et al., 2018).
- **AUTH:** took a learning to rank approach (Minas et al., 2018).
- **CNRS:** trained a logistic regression model on a large number of features over the development set (C. R. Norman et al., 2018).
- **Sheffield:** they proposed query enrichment with medical terms (Alharbi et al., 2018) .
- **UIC:** applied clustering techniques combined with an SVM classifier (Cohen & Smalheiser, 2018).
- **ECNU:** employed Paragraph2Vector to represent query and documents for similarity calculation (Wu et al., 2018).

DRAFT, to be Published in Scientometrics
https://doi.org/10.1007/s11192-020-03648-6

The results in Table 4 give more evidence that in terms of last relevant document (lastrel), our method US-RF-BioBERT is the best among all. We are also showing that this approach also reduces the workload for physicians in reviewing all the available evidence. Our method achieves the second and third places, respectively, in terms of work saved over sampling wss95 (.591) and wss100 (.627). However, *BMI combined with the knee-method* proposed by Waterloo performs better than us on wss; our method achieves the last relevant document in a lower position (2011).
Finally, for normalized cumulative gain (*ncg@k*), we reach only an eighth place in terms of ncg@20 (.402), however, we improve to 6th position in ncg@60 (.787), and we reach the sixth place in terms of norm_area (.826). As foreseen, average precision is an again metric where we do not perform very well (7th place), but we consider this a trade-off of having the best performance in terms of lastrel for a task focused on total recall.

Table 4: Benchmark of our best active learning strategies, model and document representations with other participants and baselines from CLEF eHealth 2018

model	lastrel	wss100	wss95	ncg@20	ncg@40	ncg@60	norm_area	ap
US-RF BioBERT	**2012**	.627	.591	.417	.616	.797	.826	.148
Waterloo	2655	**.756**	**.610**	**.894**	**.975**	**.996**	**.949**	.378
UNIPD	4259	.543	.396	.767	.892	.954	.896	.316
AUTH	4295	.734	.563	.881	.958	.983	.948	**.393**
CNRS	4378	.657	.510	.785	.931	.976	.928	.337
Sheffield	5519	.552	.431	.648	.866	.957	.871	.258
UIC	6385	.255	.154	.477	.633	.742	.733	.174
ECNU	7172	.029	.025	.517	.692	.764	.687	.146

7.3 Benchmark CLEF eHealth 2019

In this section, we compare the performance of our best model with other participants from CLEF eHealth 2019. In this opportunity, only three teams participated and they proposed different ranking methods, including lexical statistics for relevant term identification (Sheffield, Alharbi & Stevenson (2019)), interactive BM25 (Padua, Di Nunzio (2019)), and a combination of ranking and a "greedy" sampling strategy to estimate the number of relevant documents (ILPS, Li et al. (2019)).
It should be noted that for CLEF eHealth 2019, we only consider metrics reported in the year challenge proceedings (Kanoulas et al., 2019) since the runs are not available in the github repository as in previous years. Furthermore, we focus on Diagnostic Test Accuracy (DTA) reviews since this approach was considered in CLEF eHealth 2017 and 2018, and this way, our method is comparable amongst challenges over time.
The results in table 5 confirm that the approach we propose saves physicians workload in the task of document screening. Evidence of this is given by the

DRAFT, to be published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

lowest lastrel (786), the highest wss95 (.577) and second best wss100 (.614). On the weak side of our approach, we do not perform well in ncg metrics (*ncg@k*). We only reach the fourth position for ncg@10 (.291), ncg@20 (.502), and ncg@30 (.645) and similarly in terms of ap (.132). Overall, the results are explained by our system tendency to optimize total recall rather than ranking of the top retrieved documents. Since our main focus is aiming for total recall with the smallest effort for physicians, we do not see this as a main drawback, but of course it leaves room for improvement.

Table 5: Benchmark of our best active learning strategies, model and document representations with other participants and baselines from CLEF eHealth 2019

model	lastrel	wss95	wss100	ncg@10	ncg@20	ncg@30	ap
US-RF BioBERT	**787**	**.577**	.614	.343	.543	.693	.132
ILPS	958	.480	.526	.628	.736	.813	**.567**
Sheffield	1070	.384	.462	.404	.569	.700	.261
Padua	1111	.513	**.652**	**.630**	**.814**	**.895**	.229

8 Empirical time performance analysis

One aspect which can become an issue when using an active learning strategy is the cost of training an predicting iteratively considering an expensive learning algorithm. To study this impact, in this section we analyze empirically the time required on each active learning step, considering the labeling of ten documents per iteration, under an uncertainty sampling strategy. These experiments were run on a server with 64 GB RAM, a CPU Intel I7 with eight cores, and a SSD disk with 500GB of space. The server also has two GeForce GTX 1080 Ti GPUs, but they were not used for the active learning experiments. They were only used for obtaining BERT and BioBERT document embeddings.

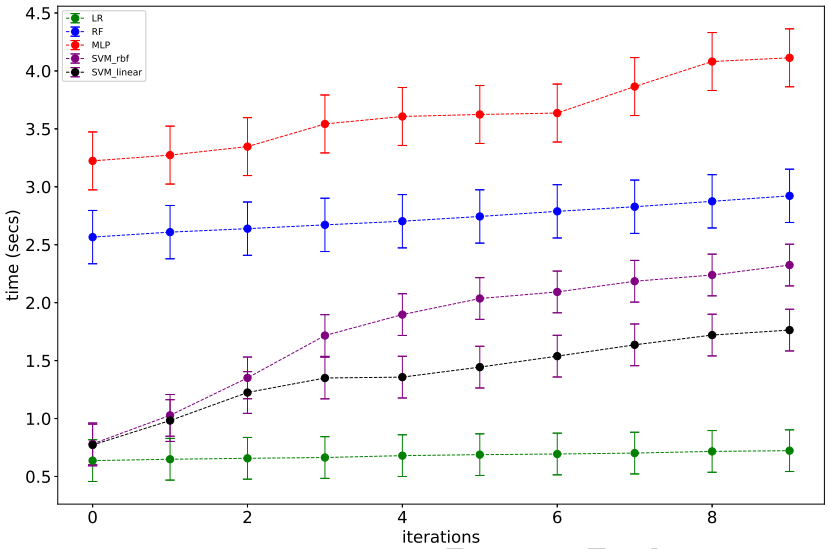
Experimental methodology. For running these experiments, we used the CLEF eHealth 2017 task 2 dataset, composed of 149,062 documents and 20 queries corresponding to the train set; and 117,557 documents on 30 questions for the test set. To represent queries and documents, we use the same setup used in section 7, where we obtained 768-dimensional text embeddings of BioBERT from the concatenation of questions and documents.

Analysis per training iteration. In the case of the training time analysis, we measured the time (seconds) it took each active learning iteration (from 0 to 9), with 10 labels by the oracle per iteration, in each of the training queries (20). We compute and plot the average training time for each model on every iteration, as well as standard deviation.

The results in Figure 14 show that Logistic Regression (LR) is the most efficient model during training. It takes less than one second per iteration and stays almost flat up to the 9th iteration, where it is training with 100 labeled documents. SVMs are almost as efficient as the LR in the first iteration, but

DRAFT, to be published in Scientometrics
https://doi.org/10.1007/s11192-020-03648-6

Fig. 14: Average training time (secs) per iteration using active learning on 20 medical questions from CLEF eHealth 2017 training dataset. Error bands show the deviation of training time among all 20 questions.



DRAFT, to be published
https://doi.org/10.1007/s11192-020-03648-6

training time grows quickly with the number of samples. Random forests are considerably less efficient than LR and SVMs in the first iterations, but the gap between RF and SVMs decreases in the latest iterations. Finally, the multi-layer perceptron, although having only two layers, takes more than 3 seconds in the initial iteration and keeps growing as the number of labeled samples increases. In practice, this means that having the oracle waiting for at least 4-5 seconds per iteration can make the process too slow to make it practical and useful.

Total training and prediction time. Concerning the analysis of training time upon the whole training dataset and prediction time among all the test datasets, we see interesting results in Table 6. For the case of training, we aggregate the seconds it took to train the model using active learning for all the questions from CLEF eHealth 2017 train dataset. (20 questions, 149,062 documents). On the other side, to obtain the prediction time, we used each of the trained models and measured the time it took to predict scores for the CLEF eHealth 2017 test set (117,557 documents).

The results on Table 6 are consistent with those in Figure 14. They show that Logistic Regression is the fastest model, taking a total time of 440 seconds, while the MLP takes up to 1467 seconds to do the complete training process for all the CLEF eHealth 2017 training set. However, a different picture is observed on prediction time, where the MLP outputs the prediction scores on the full test set (117,557 documents) in only 1.71 seconds, while the logistic regression takes 3.2 seconds, similar to the random forest model which takes

Table 6: Total training and prediction time for each machine learning model

model	training time (secs)	prediction time (secs)
RF	1467	3.40
SVM_rbf	983	6.15
SVM_linear	789	8.30
LR	440	3.20
MLP	1915	1.71

3.4 seconds. This result indicates that while training time can slow down the usefulness of these models even if they rank documents more accurately (such as RF compared to LR), after being trained they perform reasonably fast to be used in production systems, what justifies their use.

9 Discussion

In this article, we supported results from previous studies in terms of showing that active learning with an uncertainty sampling (US) strategy yields good results for the task of biomedical document screening. Our main contribution was comparing the performance of different schemes to represent documents in an active learning setting, namely TF-IDF, Word2vec, GloVe, BERT, and BioBERT. In our experiments with two datasets (CLEF eHealth and Epistemonikos), we found that an active learning strategy based on uncertainty sampling with either a BERT or BioBERT document representation, yields the best results. However, the conclusions are not completely clear in terms of the learning algorithm. In the Epistemonikos dataset, the US strategy combined with a logistic regression achieves better results in fewer iterations for retrieving documents to be labeled by an expert. Still, there are no significant differences with SVM or random forests, but LR is considerably faster for training models iteratively. In the CLEF eHealth 2017 dataset, we found that US with BioBERT document representation reaches the best performance with a random forest, leaving the logistic regression in third place after SVM. After additional analysis we found stronger similarities between the documents in train and test splits of the Epistemonikos datasets and larger differences between train and test document similiarities in the CLEF eHealth dataset, an element that might explain differences in performance of the top learning methods LR, and RF. Finally, to validate our best method (US with BioBERT + random forest), we compared our performance numbers with other participants' in CLEF eHealth 2017, 2018 and 2019 and we provided strong evidence that our method indeed saves physicians' work in the task of finding all the evidence related to a medical issue or better known as total recall problem.

Dataset complexity affects absolute metrics results. Results show that BioBERT is a robust way to represent documents for the document screening task, showing proper levels of effectiveness and efficiency for both small and large datasets. Something unusual observed in the results is that the per-

DRAFT, to be published in Scientometrics
https://doi.org/10.1007/s11192-020-03648-6

formance of the proposed framework in the CLEF eHealth dataset (e.g. best r@10=.571) is much worse than in Epistemonikos (e.g. best r@10=.717). These results could be explained by the complexity of the CLEF eHealth dataset, described by the higher density of medical terms per document compared to Epistemonikos. Nevertheless, to produce an acceptable performance in the document screening task on the CLEF eHealth dataset, there is a need to use more complex embeddings such as BERT and BioBERT. In the case of Epistemonikos, it is enough to use general embeddings such as GloVe or Word2vec to obtain good performance in this task. When comparing neural-based with traditional document representation, such as TF-IDF, there is a considerable improvement by using word and document embeddings independent of the model, active learning strategy, or dataset. The same occurred when comparing the results of active learning with relevance feedback in Rocchio and BM25, which were notoriously improved in most of the cases. Interestingly, when comparing the metric *work saved by experts*, the best results in the CLEF eHealth dataset are much better than the best results in Epistemonikos, indicating that the aforementioned dataset complexity affects the ranking of relevant documents at small cut-off positions (k=10,20), but not the ranking upon all the corpus.

Time complexity and performance trade-off. Although results show that BERT and BioBERT embeddings achieve more robust results for both small and huge datasets, there is a trade-off between time complexity and improvement in performance when using BERT or BioBERT. Indeed, when combining MLP with any of the BERT embeddings, the computation time is larger than when representing documents with Word2vec or GloVe. Another aspect worth noting is that ~~although models ranked relevant documents on the first top k, when analyzing metrics such as WSS and Lastrel%, we found that~~ even though our framework provides a considerable advance in saving work for physicians on the task of document screening, there is still room for improvement, especially for large datasets.

Comparison with other methods from CLEF eHealth. When comparing the results of our solution with the ones obtained by other methods, we showed that our approach delivers higher saves in physicians workload in the document screening task. This allows us to outstand on one of the main objectives of the evidence-based medicine discipline, which is to retrieve all the relevant evidence given a medical question. In terms of cumulative recall, our model is competitive after viewing 40 percent of the documents, but at lower levels, there is still space to improve. As expected, average precision is not competitive compared to other approaches, but this is the cost of achieving high performance on other evaluation metrics. Our method based on active learning combined with BERT embeddings surpasses traditional approaches based mostly on ranking, relevance feedback, query expansion, and classical machine learning models proposed by other participants. Interestingly, in the more current versions of the challenge, other methods did not take advantage of transformer-based embeddings, which already existed at the time. Although approaches based on relevance feedback and learning to rank algorithms yield

DRAFT, to be published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

outstanding results in recall and precision, they do not perform well on saving physicians' work, evidenced by their inferior results in metrics such as lastrel and wss.

Time complexity trade-off with model generalization. In terms of time complexity, we show that the Logistic Regression trains almost in linear time after each active learning iteration without a considerable loss in performance. We see a similar behavior while predicting, as the Logistic Regression is the second fastest model (after MLP), taking only 3.4 seconds per prediction. Based on this, for a real-world application setup, we would be recommend to use the Logistic Regression for its efficiency and competitive performance compared to other machine learning models.

SVM linear versus RBF kernel for text classification. An interesting case is the result of SVM with a linear kernel, which obtained remarkable performance in both tasks, saving work to physicians and ranking relevant documents on the top positions of the list of candidate documents. While we could expect the RBF kernel to produce better results than the linear kernel in certain cases of very small dimensionality, previous work also shows that linear kernels can produce competitive results in text classification and with considerably shorter training time Joachims (1998).

Classification models versus other factors. Overall, it could have been expected that more complex models such SVM with RBF kernel to have better performance, specially tuning their parameters, but it was not the case. The results show that simple models such as logistic regression or SVM with linear kernel were more effective than other complex ones. One possible reason for this result is that document representation and the active learning strategy are more fundamental in the learning process for this task. A similar conclusion can be reached with respect to the number of training iterations. Simple models such as logistic regression reach proper levels of effectiveness in fewer iterations than other models, for both datasets.

There are improvements over TF-IDF representation independent on the active learning strategy, machine learning model or dataset. The same occurs when using relevance feedback compared to any active learning approach for document screening task.

Future work. Based on our results and this discussion, we identify some ideas for future work. We will test other paradigms for more scalable learning, such as weak supervision, where using vast amounts of weak labeled data sources improve the model capability to generalize to new cases (Dehghani et al., 2017) requiring even fewer expert labels for training a classification models (Ratner et al., 2017).

Concerning embeddings, we will test different values of sensitive parameters, such as term normalization and choice of training collection (Roy et al., 2018). Although the focus of the current work is to show the impact of language models on active learning strategies for medical literature, it would be important to test the best active learning strategy with actual users. For this, it is necessary to integrate our model into an interface and conduct a user evaluation with actual physicians. For future work, we plan to integrate our best model

DRAFT, to be published in Scientometrics
https://doi.org/10.1007/s11192-020-03648-6

with the Epistaid interface developed by Donoso-Guzmán & Parra (2018), and then design and conduct a user study.

Limitations. We have identified some limitations which we describe here. For active learning, there are several additional sampling strategies, such as query-by-committee and gradient-based methods (Settles, 2012). We did not experiment with them due to its computational complexity: in practice, physicians would need to wait several minutes between iterations of labeling, which in a real setting is not feasible.

Another important limitation is that although we outperformed several methods in our CLEF eHealth 2017, 2018 and 2019 evaluation, the comparisons are not entirely fair, since we had no limits in the number of runs, we had the advantage to analyze the approaches followed by the other participants.

10 Summary and Conclusion

In this article we studied how recent document representations based on neural language models such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), BERT (Devlin et al., 2018), and BioBERT (J. Lee et al., 2019) compare on a task of biomedical document screening in an active learning setting. Our goal was to determine the best combination of document representation, learning method, and sampling strategy to help physicians reducing their manual work while screening biomedical documents in the practice of Evidence-based Medicine. Our analyses, conducted on two datasets (CLEF eHealth and Epistemonikos), indicate that recent full-text models such as BERT or BioBert, provide better performance. Yet, this difference was significant only in the smaller but more challenging dataset CLEF eHealth. Our analyses indicate that ranking gets increasingly difficult in a dataset with a more specialized vocabulary and with small term overlap between query (clinical question) and potential relevant documents. On the larger Epistemonikos dataset, the difference in performance between the word-embedding model Word2Vec was not significant different compared to BERT. In contrast, the cost of training these full-text embedding models is way higher, since they are based on the transformer architecture and require specialized hardware (TPU vs. GPU) to be trained more efficiently. Another interesting but expected result is the improvement over relevance feedback and TF-IDF, which is considerable when using an active learning framework. Another interesting result shows that logistic regression (as well SVM with linear kernel and random forests) perform quite well with uncertainty-based sampling compared to more sophisticated learning models such as RBF-kernel SVM. Furthermore, logistic regression has the additional benefit of converging very quickly compared to other learning algorithms and sampling strategies. We then recommend using logistic regression or random forests with uncertainty sampling strategy combination for performing active learning for biomedical document screening, since the key in performance seems to be more related to the document representation.

DRAFT to be published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

Moreover, we provided evidence that more complex embeddings (BERT or BioBERT) had a better performance on the CLEF eHealth dataset, which is harder than the Epistemonikos dataset, explained by the higher proportion of medical terms present and the smaller overlap between queries and relevant documents.
To further support the performance of our method, we compared it against other competing approaches on CLEF eHealth challenges 2017, 2018, and 2019. We showed better results in terms of work saved by physicians by using our active learning framework, which was one of our main objectives. In terms of cumulative gain, our best model can retrieve a considerable proportion of relevant documents after reviewing only 40 percent of the retrieved documents.
Finally, regarding the time complexity, there is a trade-off between choosing a model that has better performance but takes more time per active learning re-training iteration (i.e., Random Forest), or choosing a model that takes little time to train and has moderately competitive results, as the logistic regression. Our suggestion is that logistic regression has a better general value in terms of predictive ranking accuracy, as well as training and prediction time cost.
For future work, we consider using a richer representation of the documents by adding other features, such as author information, year of publication, as well as images represented as neural visual embeddings using AlexNet, VGG, or ResNet. We will also study other learning strategies beyond active learning, such as weak supervision (Ratner et al., 2017). Another potential improvement would be to experiment with representations of more structured descriptions of the documents, which are specified within some abstracts, such as detailed description of objective, methodology, materials and implications. Lastly, it would be interesting to test our best active learning strategy in an interface and conduct a user study for validating the practical validity of our research.

11 Acknowledgments

This research was funded by ANID Chile, Fondecyt Grant 1191791, and by the Millenium Institute Foundational Research on Data (IMFD).

12 References

Adeva, J. G., Atxa, J. P., Carrillo, M. U., & Zengotitabengoa, E. A. (2014). Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications*, *41*(4), 1498–1508.

Alharbi, A., Briggs, W., & Stevenson, M. (2018). Retrieving and ranking studies for systematic reviews: University of sheffield's approach to clef ehealth 2018 task 2. In *Ceur workshop proceedings* (Vol. 2125).

Alharbi, A., & Stevenson, M. (2017). Ranking abstracts to identify relevant evidence for systematic reviews: The university of sheffield's approach to clef ehealth 2017 task 2. In *Clef (working notes)*.

DRAFT, to be published in Scientometrics
https://doi.org/10.1007/s11192-020-03648-6

Alharbi, A., & Stevenson, M. (2019). Ranking studies for systematic reviews using query adaptation: University of sheffield's approach to clef ehealth 2019 task 2. In *Clef (working notes)*.

Bekhuis, T., Tseytlin, E., Mitchell, K. J., & Demner-Fushman, D. (2014). Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence. *PloS one*, *9*(1), e86277.

Chen, J., Chen, S., Song, Y., Liu, H., Wang, Y., Hu, Q., ... Yang, Y. (2017). Ecnu at 2017 ehealth task 2: Technologically assisted reviews in empirical medicine. In *Clef (working notes)*.

Choi, S., Ryu, B., Yoo, S., & Choi, J. (2012). Combining relevancy and methodological quality into a single ranking for evidence-based medicine. *Information Sciences*, *214*, 76–90.

Cohen, A. M., & Smalheiser, N. R. (2018). Uic/ohsu clef 2018 task 2 diagnostic test accuracy ranking using publication type cluster similarity measures. In *Ceur workshop proceedings* (Vol. 2125).

Cormack, G. V., & Grossman, M. R. (2016). " when to stop" waterloo (cormack) participation in the trec 2016 total recall track. In *Trec*.

Cormack, G. V., & Grossman, M. R. (2017). Technology-assisted review in empirical medicine: Waterloo participation in clef ehealth 2017. In *Clef (working notes)*.

Dehghani, M., Zamani, H., Severyn, A., Kamps, J., & Croft, W. B. (2017). Neural ranking models with weak supervision. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval* (pp. 65–74).

Del Fiol, G., Michelson, M., Iorio, A., Cotoi, C., & Haynes, R. B. (2018, Jun 25). A deep learning method to automatically identify reports of scientifically rigorous clinical research from the biomedical literature: Comparative analytic study. *J Med Internet Res*, *20*(6).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Di Nunzio, G. M. (2019). A distributed effort approach for systematic reviews. ims unipd at clef 2019 ehealth task 2. In *Clef (working notes)*.

Di Nunzio, G. M., Beghini, F., Vezzani, F., & Henrot, G. (2017). An interactive two-dimensional approach to query aspects rewriting in systematic reviews. ims unipd at clef ehealth task 2. In *Clef (working notes)*.

Di Nunzio, G. M., Ciuffreda, G., & Vezzani, F. (2018). Interactive sampling for systematic reviews. ims unipd at clef 2018 ehealth task 2. In *Clef (working notes)*.

Donoso-Guzmán, I., & Parra, D. (2018). An interactive relevance feedback interface for evidence-based health care. In *23rd international conference on intelligent user interfaces* (pp. 103–114).

Elliott, J. H., Turner, T., Clavisi, O., Thomas, J., Higgins, J. P., Mavergames, C., & Gruen, R. L. (2014). Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS medicine*, *11*(2), e1001603.

DRAFT. To be published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

Figueroa, R. L., Zeng-Treitler, Q., Ngo, L. H., Goryachev, S., & Wiechmann, E. P. (2012). Active learning for clinical text classification: is it better than random sampling? *Journal of the American Medical Informatics Association*, *19*(5), 809–816.

Goeuriot, L., Kelly, L., Suominen, H., Névéol, A., Robert, A., Kanoulas, E., ... Zuccon, G. (2017). Clef 2017 ehealth evaluation lab overview. In *International conference of the cross-language evaluation forum for european languages* (pp. 291–303).

Goodwin, T. R., & Harabagiu, S. M. (2018). Knowledge representations and inference techniques for medical question answering. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *9*(2), 14.

Grossman, M. R., Cormack, G. V., & Roegiest, A. (2016). Trec 2016 total recall track overview. In *Trec.*

Hashimoto, K., Kontonatsios, G., Miwa, M., & Ananiadou, S. (2016). Topic detection using paragraph vectors to support active learning in systematic reviews. *Journal of biomedical informatics*, *62*, 59–65.

Hollmann, N., & Eickhoff, C. (2017). Ranking and feedback-based stopping for recall-centric document retrieval. In *Clef (working notes).*

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Hughes, M., Li, I., Kotoulas, S., & Suzumura, T. (2017). Medical text classification using convolutional neural networks. *Stud Health Technol Inform*, *235*, 246–50.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning* (pp. 137–142).

Kalphov, V., Georgiadis, G., & Azzopardi, L. (2017). Sis at clef 2017 ehealth tar task. In *Ceur workshop proceedings* (Vol. 1866, pp. 1–5).

Kanoulas, E., Li, D., Azzopardi, L., & Spijker, R. (2017). Clef 2017 technologically assisted reviews in empirical medicine overview. In *Ceur workshop proceedings* (Vol. 1866, pp. 1–29).

Kanoulas, E., Li, D., Azzopardi, L., & Spijker, R. (2018). Clef 2018 technologically assisted reviews in empirical medicine overview. In *Ceur workshop proceedings* (Vol. 1866, pp. 1–34).

Kanoulas, E., Li, D., Azzopardi, L., & Spijker, R. (2019). Clef 2019 technology assisted reviews in empirical medicine overview. In *Ceur workshop proceedings* (Vol. 2380).

Keselman, A., & Smith, C. A. (2012). A classification of errors in lay comprehension of medical documents. *Journal of biomedical informatics*, *45*(6), 1151–1163.

Lagopoulos, A., Anagnostou, A., Minas, A., & Tsoumakas, G. (2018). Learning-to-rank and relevance feedback for literature appraisal in empirical medicine. In *International conference of the cross-language evaluation forum for european languages* (pp. 52–63).

Lee, G. E. (2017). A study of convolutional neural networks for clinical document classification in systematic reviews: sysreview at clef ehealth

DRAFT, to be published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

2017.

Lee, G. E., & Sun, A. (2018). Seed-driven document ranking for systematic reviews in evidence-based medicine. In *The 41st international acm sigir conference on research & development in information retrieval* (pp. 455–464).

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Li, D., Kanoulas, E., et al. (2019). Automatic thresholding by sampling documents and estimating recall: Ilps@ uva at tar task 2.2. In *Ceur workshop proceedings* (Vol. 2380).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Minas, A., Lagopoulos, A., & Tsoumakas, G. (2018). Aristotle university's approach to the technologically assisted reviews in empirical medicine task of the 2018 clef ehealth lab. In *Clef (working notes)*.

Miwa, M., Thomas, J., O'Mara-Eves, A., & Ananiadou, S. (2014). Reducing systematic review workload through certainty-based screening. *Journal of biomedical informatics*, *51*, 242–253.

Mo, Y., Kontonatsios, G., & Ananiadou, S. (2015). Supporting systematic reviews using lda-based document representations. *Systematic reviews*, *4*(1), 172.

Nogueira, R., Yang, W., Cho, K., & Lin, J. (2019). Multi-stage document ranking with bert. *arXiv preprint arXiv:1910.14424*.

Norman, C., Leeflang, M., & Névéol, A. (2017). Limsi@ clef ehealth 2017 task 2: Logistic regression for automatic article ranking.

Norman, C. R., Leeflang, M. M., & Névéol, A. (2018). Limsi@ clef ehealth 2018 task 2: Technology assisted reviews by stacking active and static learning. *CLEF (Working Notes)*, *2125*, 1–13.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Qiao, Y., Xiong, C., Liu, Z., & Liu, Z. (2019). Understanding the behaviors of bert in ranking. *arXiv preprint arXiv:1904.07531*.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, *11*(3), 269–282.

Roy, D., Ganguly, D., Bhatia, S., Bedathur, S., & Mitra, M. (2018). Using word embeddings for information retrieval: How collection and term

DRAFT. To be published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

normalization choices affect performance. In *Proceedings of the 27th acm international conference on information and knowledge management* (pp. 1835–1838).

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Scells, H., Zuccon, G., Deacon, A., & Koopman, B. (2017). Qut ielab at clef ehealth 2017 technology assisted reviews track: initial experiments with learning to rank. In *Ceur workshop proceedings: Working notes of clef 2017: Conference and labs of the evaluation forum* (Vol. 1866, pp. Paper–98).

Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *6*(1), 1–114.

Singh, G., Marshall, I., Thomas, J., & Wallace, B. (2017). Identifying diagnostic test accuracy publications using a deep model. In *Ceur workshop proceedings* (Vol. 1866).

Singh, J., & Thomas, L. (2017). Iiit-h at clef ehealth 2017 task 2: Technologically assisted reviews in empirical medicine. In *Clef (working notes)*.

van Altena, A. J., & Olabarriaga, S. D. (2017). Predicting publication inclusion for diagnostic accuracy test reviews using random forests and topic modelling. In *Clef (working notes)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., Schmid, C. H., Bertram, L., ... Trikalinos, T. A. (2012). Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining. *Genetics in medicine*, *14*(7), 663.

Wallace, B. C., Small, K., Brodley, C. E., & Trikalinos, T. A. (2010). Active learning for biomedical citation screening. In *Proceedings of the 16th acm sigkdd international conference on knowledge discovery and data mining* (pp. 173–182).

Wu, H., Wang, T., Chen, J., Chen, S., Hu, Q., & He, L. (2018). Ecnu at 2018 ehealth task 2: Technologically assisted reviews in empirical medicine. *Methods*, *4*(5), 7.

Yang, Y.-Y., Lee, S.-C., Chung, Y.-A., Wu, T.-E., Chen, S.-A., & Lin, H.-T. (2017). *libact: Pool-based active learning in python.*

Yu, Z., & Menzies, T. (2017). Data balancing for technologically assisted reviews: Undersampling or reweighting. In *Clef (working notes)*.

DRAFT to be published in Scientometrics
https://doi.org/10.1007/s11192-020-03648-6

Appendix

Appendix I: Leave One Out cross-validation CLEF eHealth 2017

In this section we do a leave one out cross-validation evaluation. Concerning model input we use the same method to represent documents and medical

questions pairs representation proposed in section 7, allowing it to be possible to use this evaluation methodology. Now, instead of having one model for each medical question, we train a general model to make predictions for new questions and then rank the documents. We evaluate the performance of our active learning framework using leave one out cross-validation for the 50 queries included in the CLEF eHealth 2017 task 2 dataset on our four best combinations of active learning strategies, machine learning model, and language model representations (US-BioBERT-RF, US-BERT-RF, US-BioBERT-LR, and US-BERT-LR). The evaluation metrics used were the average of each medical question testing and training with the rest on the last relevant document found (lastrel), work saved oversampling (wss95 and wss100), average precision (ap), and normalized cumulative gain at recall@k% (20,40,60).

Table 7: Active Learning on CLEF eHealth 2017 dataset using leave one out cross-validation using US-BioBERT-RF, US-BioBERT-LR, US-BERT-RF and US-BERT-LR. Results show the average metrics for training on the whole dataset except a given queries and testing on each of the 50 queries from CLEF eHealth 2017 and the standard error.

Model	lastrel	wss100	wss95	ncg@20	ncg@40	ncg@60	ap
US-BioBERT-RF	968±132	.603±.05	.624±.040	.652±.049	.827±.36	.925±.025	.182±.028
US-BERT-RF	1079±198	.565±.051	.587±.047	.610±.050	.772±.039	.870±.026	.153±.029
US-BioBERT-LR	1453±262	.486±.054	.511±.052	.538±.051	.723±.041	.834±.029	.088±.031
US-BERT-LR	1707±285	.424±.054	.453±.051	.481±.049	.679±.041	.809±.029	.072±.027

For carrying out this evaluation, we compare the leave one out cross-validation on our best four models obtained from experiments made in section 6 for the CLEF eHealth 2017 task 2 dataset. We train with all the questions and test with one, repeating this process iteratively until we have evaluation results for each medical query. Regarding the metrics used for the evaluation, we used parameters related to saved work (wss100 and wss95), accumulated recall (ncg20,ncg40, and ncg60), and precision (ap).

The results on table 7 show that in terms of the last relevant document (lastrel), the method US-RF-BIOBERT is the best among our other models. With these results, we confirm that this is our best model for the task of retrieving all the relevant evidence given a medical question. More evidence of this effect is provided by work-saved oversampling (wss100 and wss95), which indicates that our best approach allows the physician to save near 60% of their work. Then, concerning cumulative gain (ncg20, ncg40, ncg60), US-BioBERT-RF ranks a 92% of the relevant documents on the first 60% of the candidates list. Finally, in terms of precision, US-RF-BIOBERT obtains the best results among the other models; however, it finds only 18.2 percent of the relevant documents over the total of retrieved ones.

DRAFT, to be published in Scientometrics https://doi.org/10.1007/s11192-020-03648-6

Appendix II: CLEF eHealth 2017 task 2 Test Dataset

Table 8: Distribution of relevant and total documents in CLEF eHealth Test Dataset

topic id	relevant_documents	total_documents	% relevants
CD010409	492	13283	3.70
CD010339	402	11653	3.45
CD011548	390	11623	3.36
CD011549	138	11622	1.19
CD010783	105	6905	1.52
CD011145	99	6894	1.44
CD011975	82	6636	1.24
CD011984	66	6633	1.00
CD009925	65	5688	1.14
CD008643	54	5252	1.03
CD009593	43	4537	0.95
CD012019	42	4352	0.97
CD008782	41	3706	1.11
CD009591	41	3682	1.11
CD010276	38	2658	1.43
CD010173	37	2658	1.39
CD010653	33	2429	1.36
CD009519	30	1915	1.57
CD009323	27	1732	1.56
CD008803	25	1612	1.55
CD008686	24	1586	1.51
CD007431	24	1015	2.36
CD008054	22	989	2.22
CD010438	21	977	2.15
CD009647	21	849	2.47
CD007394	21	848	2.48
CD009786	20	741	2.70
CD010633	18	688	2.62
CD010632	16	682	2.35
CD009372	15	679	2.21
CD009551	13	653	1.99
CD011134	12	612	1.96
CD009020	11	568	1.94
CD007427	10	553	1.81
CD009185	10	488	2.05
CD008691	7	396	1.77
CD009944	7	394	1.78
CD008081	6	332	1.81
CD010386	6	318	1.89
CD010023	3	297	1.01
CD010771	3	278	1.08
CD010772	3	276	1.09
CD009135	3	250	1.20
CD010896	3	108	2.78
CD010542	3	104	2.88
CD010775	1	102	0.98
CD010705	1	56	1.79
CD010860	1	40	2.50
CD008760	1	19	5.20

DRAFT, to be published in Scientometrics
https://doi.org/10.1007/s11192-020-03648-6

Appendix III: Epistemonikos dataset

Table 9: Distribution of a sample of twenty questions with their relevant and total documents in Epistemonikos Test Dataset

topic id	total_documents	relevant_documents	% relevants
537f18c45f844e05567299be	480	24	5.00
562eb0c1d8307f0e503e903f	180	9	5.00
55f60bf3dfbaca256b4a71f2	648	32	4.94
57728ceadfbaca1d9ef85103	936	46	4.91
585931b7d8307f1d1b240118	174	8	4.60
5835043bdfbaca2943ba929d	192	9	4.69
5215526ea2e3a9261f61a72a	1926	96	4.98
56a52e50dfbaca0cb759e8be	282	14	4.96
521c78e3659e937c5c4a79c2	150	7	4.67
52b49e05659e933d89c686fa	1428	71	4.97
55dd00ed18d84e79307a9eeb	906	45	4.97
55b6e0d0dfbaca5792d299cb	216	10	4.63
51a88cbbd5d70f4d37c6ba1f	204	10	4.90
53ae4b645f844e42cec766d8	168	8	4.76
5213ead3a2e3a92621390d5c	288	14	4.86
5535965b18d84e32c05ed775	144	7	4.86
56aa88fdd8307f1c69e69856	168	8	4.76
53e61aea5f844e65f954d2cb	984	49	4.98
5627e644d8307f6d64b8d2ac	402	20	4.98
578ea553dfbaca32c9cdcbd8	396	19	4.80

DRAFT, to be published in Scientometrics
https://doi.org/10.1007/s11192-020-03648-6