UWaterlooMDS at the TREC 2021 Health Misinformation Track

MUSTAFA ABUALSAUD[1], IRENE XIANGYI CHEN[2], KAMYAR GHAJAR[1], LINH NHI PHAN MINH[1], MARK D. SMUCKER[3], AMIR VAKILI TAHAMI[3], DAKE ZHANG[1],

[1] David R. Cheriton School of Computer Science, University of Waterloo

[2] Department of Electrical and Computer Engineering, University of Waterloo

[3] Department of Management Sciences, University of Waterloo

In this report, we discuss the experiments we conducted for the TREC 2021 Health Misinformation Track. For our manual runs, we used an improved version of our high-recall retrieval system [2] to manually search and judge documents. The system is built to efficiently retrieve the most-likely relevant documents based on a Continuous Active Learning (CAL) model and allows a speedy document assessment phase. Using the judged documents, we built CAL models to score documents that are part of our filtered collections. We also experimented with neural reranking methods based on question answering and stance detection methods to modify our CAL-based runs and a traditional BM25 run. For our automatic runs, we filtered the collection by running PageRank with a seed set of reliable domains and then using a text classifier and further refined the collection by including only medical web pages. We then ran traditional BM25 on this smaller and more reliable collection.

1 INTRODUCTION

The task of the 2021 Health Misinformation Track is an ad-hoc retrieval task where researchers design and build retrieval technologies to retrieve credible and correct information while avoiding non-credible and incorrect information in order to help search users make correct decisions about their health concerns.

We submitted both automatic and manual runs. Most of these runs were constructed using one of three filtered collections that are subsets of the track's collection. The motivation behind the filtering was to produce a collection with high-quality health-related documents so that our retrieval methods would be able to find correct and credible documents and avoid retrieving incorrect and low-credibility documents. This effectively reduced the size of the given collection, which also allowed for faster data processing and retrieval. Different techniques were used to filter the collections resulting in collections of different sizes.

We used several methods to construct our manual runs. The first method utilized our high-recall retrieval system [2], which is based on Continuous Active Learning (CAL), to score documents with assessors making manual judgements for the track's topics. The second method implemented a combination of CAL, and the RoBERTa language model [6], where we scored paragraphs using CAL trained on assessors' manual judgements and then reranked based on RoBERTa to match each topic has given stance field. The last method was to fine-tune T5-Large [10] to acquire a binary classification model to predict the stance of each document. We built our automatic runs using Anserini's BM25 on our different filtered collections.

Results show that in terms of the compatibility measure, using our filtered collections produced runs with better performances than just using the entire collection (as was done to create the baseline run). Based on the nDCG measure, several of our runs achieved higher scores than the baseline run. Precision at 10 (P@10) scores show that the use of our filtered collections produced runs with better credibility. Overall, creating filtered collections allowed for a boost in performance.

Table 1. Summary of our filtered collections used to build our runs.

Tag	Name	# Documents	Description
A	`c4/en.noclean`	1,063,805,381	The track's collection. The collection is comprised of text extracts from the April 2019 snapshot of Common Crawl web crawl corpus.
M	Reliable Medical Collection	3,568,939	This collection only includes documents with domains having an HONcode certification (see www.hon.ch for more details) or are part of 13 handpicked health related websites (e.g. kidshealth.org).
C	**M** + 10k BM25	4,040,012	This collection expands **M** by including the top 10k BM25 results per topic from the complete collection.
E	Expanded Reliable Medical Collection	1,829,111	This collection expands **M** by adding domains that were linked them. We select the top 10,000 domains and filter out all non-medical documents with a medical text classifier, thus ending up with a smaller collection than **M**.
T	3k BM25	144,367	This collection contains the top 3k BM25 results per topic from the complete collection.

2 COLLECTION

2.1 Original collection (**A**)

This year, the track used the `c4/en.noclean` version of the c4 dataset[1]. The collection is comprised of text extracts from the April 2019 snapshot of Common Crawl web corpus and contains about 1 billion English documents. The compressed size of the collection is 2.2 terabytes.

2.2 Filtered collections

We created different subsets of the collection that focus on filtering out non-medical and unreliable documents. These subsets are much smaller than the original collection and allow easier and faster data processing and retrieval. Most of our runs were constructed using one of these filtered collections. Table 1 shows a brief summary of each collection.

*2.2.1 Reliable Medical Collection (**M**).* In this collection, we focused on filtering out non-credible and non-health-related websites. Our filtering method is based on detecting domains with HONCode certification[2]. The HONcode certification is maintained by a non-profit and non-governmental organization named The Health On the Net Foundation (HON) and is created to promote access to useful and reliable health information online. The certification assessment is carried out by medical experts and is only given to websites offering health information that is deemed reliable. Such websites include the World Health Organization[3] and Mayo Clinic[4], among others. According to the foundation's website, there are currently more than 8,000 websites that have been certified.

To the best of our knowledge, the list of HONCode certified domains are not publicly available. In our case, we used the HON foundation's browser plugin[5] that was designed for people to easily identify HONCode certified websites while browsing and searching the internet. The extension works by matching the MD5 hash of a website

[1] https://www.tensorflow.org/datasets/catalog/c4
[2] https://www.hon.ch/en/certification.html
[3] https://www.who.int
[4] https://www.mayoclinic.org/
[5] https://github.com/healthonnet/hon-honcode-extension

domain to a public list of HONCode certified MD5 domain hashes[6] that the plugin uses. To extract the HONCode domains in our collection, we iterated and calculated MD5 hashes for every unique domain in the collection and recorded domains that are part of the list. In total, we found 2094 domains. We excluded domains that are part of `wordpress.com` or `blogpost.com`. In addition to these domains, we added 13 health-related websites that we believe are reliable but are not HONCode certified (e.g., kidshealth.org, aarp.org, etc.). Some of these websites were manually selected from Alexa Global Traffic Rank. With this new set of domains, we constructed the collection by including only documents from domains that are part of our set, yielding a total of 3,568,939 documents.

*2.2.2 Reliable Medical Websites + 10k BM25 Collection (**C**).* While the previous collection (**M**) focuses on reliable health-related websites, our HONCode filtering method filters out documents that are potentially very useful and relevant to the track's topics. Such documents can be from websites that are deemed credible but are not HONCode certified, non-health-related websites that include correct health information, or websites that their credibility is unknown but provide correct health information. In this collection, we expand on our previous collection by including the top 10k BM25 search results for each topic retrieved from the original collection. We used Anserini's[7] BM25 with its default parameters to retrieve the top 10k results. The additional 10k results are used to try to include as many relevant documents as possible for each topic, regardless of their credibility or domain focus (e.g., news websites). In total, this collection has 4,040,012 documents.

*2.2.3 Expanded Reliable Medical Collection (**E**).* The purpose of this collection is to explore an alternative query independent method of expanding **M** while keeping the document count low. Using the hosts in M as a base, we expanded this list using the common crawl host graph [8]. The hosts graph contains roughly 4 million nodes and 4 billion edges. This expanded collection contains domains for reputable journals and organizations not included in **M** (e.g. bmj.com) but also various irrelevant and/or non-reputable domains (e.g. blog posts). The number of documents for these hosts also greatly increased. We used two steps to filter this expanded collection.

In our first step, we aim to expand the list of reliable medical websites in **M**. We do this by calculating PageRank scores in a subset of the common crawl host-level graph. The subset is created as follows: For the nodes, we take the domains in **M** and all the domains they link to. For the edges, we take all the edges with a **M** domain as its source. We calculate PageRank but only jump randomly to domains in **M**. This approach is similar to Topic Sensitive PageRank [5]. After calculating each domain's score, we take the top 10,000 domains. With this approach, we end up with roughly 30 million documents. Many of these documents are non-medical and irrelevant to the task at hand. Thus, in the next step, we aim to filter these out from the collection.

In our next step, we filter out these non-medical pages with a rudimentary medical text classifier. For the positive samples in our training data, we use the top 100 documents retrieved from collection **A** by Anserini's BM25 using the 2019 track topics as queries. The idea is that since these queries are medical in nature and the top 100 documents are pseudo-relevant, they would represent a good mix of medical pages present in the collection. For our negative samples, we sample randomly from $A - M$. We train a model using a linear support vector machine and validate with 5-fold cross-validation where we train on 40 topics and test with 11 in each fold. With this classifier, we filter out documents whose text is classified as non-medical and are left with 1,829,111 documents.

[6] https://www.honcode.ch/HONcode/Plugin/listeMD5.txt
[7] https://github.com/castorini/anserini
[8] https://commoncrawl.org/2020/02/host-and-domain-level-web-graphs-novdecjan-2019-2020/

Table 2. List of submitted runs and their details.

	Run tag	Collection	Manual Assessment	Fields					Ranking method(s)
				query	desc.	stance	narr.	evidence	
Automatic	baselineBM25	A		•					BM25 (Anserini)
	WatSAM-BM25	M		•					
	WatSAE-BM25	E		•					
	WatSAE-BM25RM3	E		•					
	WatSAE-BM25-RR	E		•					
Manual	WatSMM-CAL	M	✓	•	•				CAL
	WatSMM-CALHC	M	✓	•	•				CAL
	WatSMM-CALPR	M	✓	•	•				CAL
	WatSMM-Fused	M	✓	•	•				CAL
	WatSMC-CAL	C	✓	•	•				CAL
	WatSMM-CALQA100	M	✓	•	•	•			CAL + RoBERTa
	WatSMM-CALQAAll	M	✓	•	•	•			CAL + RoBERTa
	WatSMC-CALQA100	C	✓	•	•	•			CAL + RoBERTa
	WatSMC-CALQAAll	C	✓	•	•	•			CAL + RoBERTa
	WatSMC-CALQAHC1	C	✓	•	•	•			CAL + RoBERTa
	WatSMC-CALQAHC2	C	✓	•	•	•			CAL + RoBERTa
	WatSMT-SD-S1	T		•		•			BM25 (Anserini) + T5
	WatSMT-SD-S2	T		•		•			BM25 (Anserini) + T5
	WatSMC-Correct	C	✓	•	•	•	•	•	CAL

3 SUBMITTED RUNS

Table 2 shows the list of runs we submitted to the track. In total, we submitted 19 runs — 5 automatic and 14 manual runs. We describe the details of each run below.

3.1 Automatic Runs

3.1.1 `baselineBM25`. As a baseline for the other runs, we have retrieved the top 1000 documents per topic using Anserini's implementation of BM25 scoring with their default parameters of $k_1 = 0.9$ and $b = 0.4$. We built the index for this run from our **A** collection (original collection) mentioned in section 2.1. For indexing, we used Anserini's `IndexCollection` program with its default English analyzer that uses Apache Lucene's(v8.0) implementations of the standard tokenizer, Porter stemmer and some typical text cleansing techniques such as stopwords filtering and lowercase conversions for English documents' text analysis.

3.1.2 `WatSAM-BM25`. In this run, we used Anserini's BM25 on our **M** collection that contains reliable medical websites only. We used Anserini's default English analyzer for indexing and retrieval with the same process as the `baselineBM25` run, but using our **M** collection.

3.1.3 `WatSAE-BM25`. In this run we use Anserini's BM25 implementation on the **E** collection explained in section 2.2.3. We used Anserini's default English analyzer for indexing and default parameters for retrieval in this run as well.

3.1.4 `WatSAE-BM25RM3`. In this run we use Anserini's BM25 + RM3 implementation on the **E** collection explained in section 2.2.3. We used Anserini's default parameters.

3.1.5 `WatSAE-BM25-RR`. In this run, we rerank the top 1000 results returned by WatSAE-BM25 with a series of features. These features are as follows:

- Presence of specific words in the URL such as "https","buy", "shop" and "product".
- Ratio of medical to total documents on the domain
- PageRank score of the domain in the hosts' web graph
- PageRank score of domain in graph subset discussed in section 2.2.3
- whether the domain is part of the Reliable Medical Collection

The features were normalized using Z-score normalization. We use the sigmoid function to map the features into the [0,1] range. We give each feature a weight that we fine-tuned using a small number of manual judgements for the 2019 track topics. We calculate the new scores as follows:

$$\text{Score}(q,d) = \text{BM25}(q,d) + \sum_{i}^{n} \lambda_i \text{sigmoid}\left(F_i(d)\right)$$

where λ_i) is the weight hyper-parameter for the ith feature, $F_i(d)$ is the normalized score of document d for the ith feature.

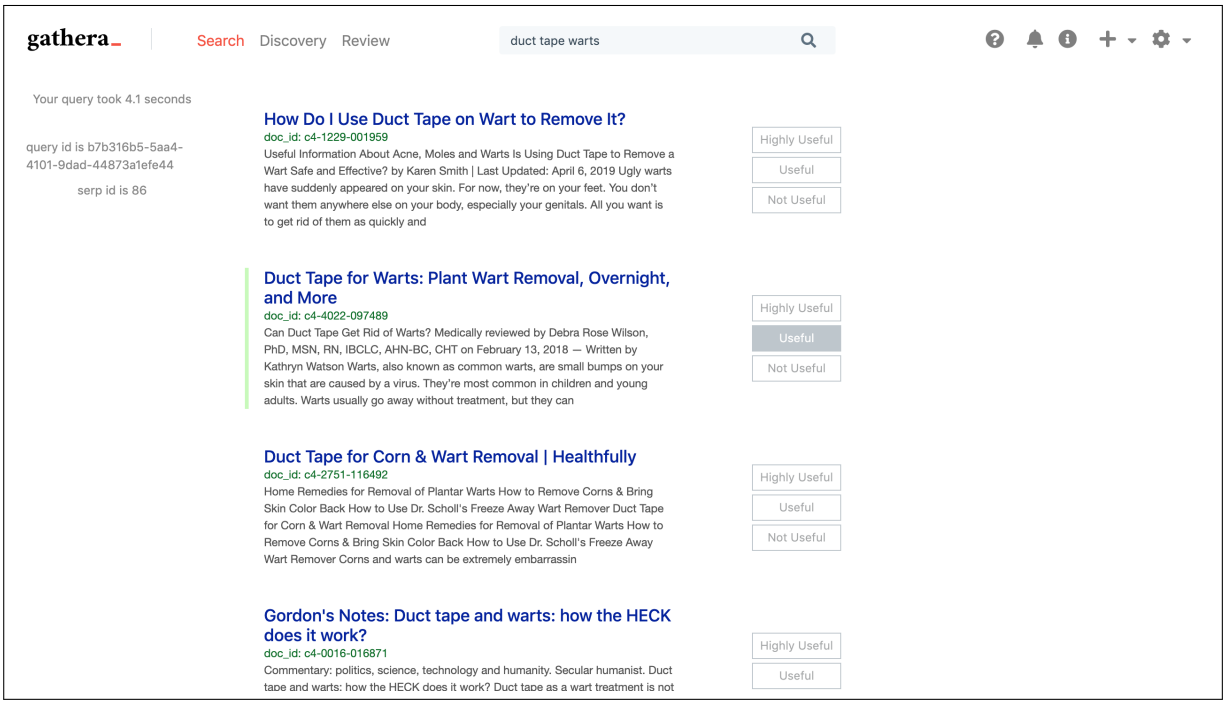
3.2 Manual Runs

Generally, our manual runs can be classified into three main categories: CAL-based active learning, stance detection model-based reranking, and CAL-assisted human assessments.

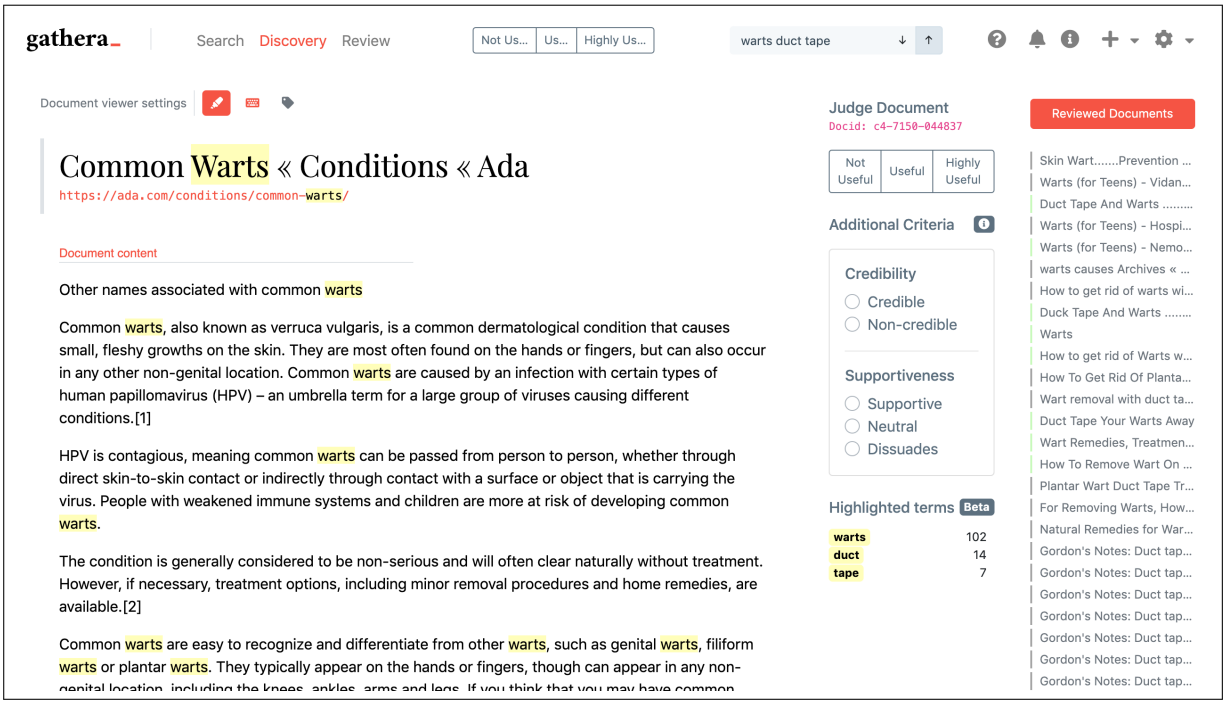
3.2.1 CAL-based Manual Assessments. We used an improved version of our high-recall retrieval system [2] to manually assess some of the documents in our filtered collections[9]. This system was successfully used in previous TREC tracks for building manual runs [1, 11]. The system has two main components: "Search", which allows interactive search and judging, and "Discovery", which is based on Continuous Active Learning (CAL) that prompts the user with the next-likely relevant documents based on a logistic regression model. Given a seed query or seed documents, the system initiates a CAL session and presents the user with documents that are predicted as most-likely relevant to the seed. As the user judges a document, the model is retrained with the new judgment, and the next likely relevant document is then presented to the user. This process keeps going until all documents are judged or when the user reaches their allocated judging budget for the topic. In our case, we allocated fixed time periods of 10 or 5 minutes where the user keeps judging.

Screenshots of both components are shown in Figure 1. The "Search" interface in Figure 1a uses Anserini's implementation of BM25 scoring to retrieve documents for user-submitted queries. Any judgment made while in the "Search" interface is used to retrain the CAL model. Figure 1b shows an example of the "Discovery" interface, where the next-likely relevant document from the CAL model is presented to the user after having previously judged some documents. As the interface shows, the system allows different judging criteria to be submitted for each document (e.g., usefulness, credibility, and supportiveness). For all our CAL-based runs except `WatSMC-Correct`, only the usefulness of a document is used for training, where highly useful and useful documents are treated as positive samples and not useful documents as negative samples. To speed up the judging process in our document assessing phase of building the runs, we only focused on judging the usefulness of documents and left other criteria unjudged.

[9] https://github.com/UWaterlooIR/gathera



(a) Search interface.



(b) Discovery interface.

Fig. 1. Screenshots of our document assessment system.

Table 3. Top 10 ranked domains by harmonic centrality and PageRank scores in the **M** collection.

Table 4. Host names sorted by harmonic centrality score.

Rank	Score	Host name
1	1.000	psychologytoday.com
2	0.997	webmd.com
3	0.978	fda.gov
4	0.955	healthline.com
5	0.951	columbia.edu
6	0.948	heart.org
7	0.946	mayoclinic.org
8	0.946	cancer.org
9	0.941	psychcentral.com
10	0.939	aarp.org

Table 5. Host names sorted by PageRank score.

Rank	Score	Host name
1	1.000	fda.gov
2	0.763	webmd.com
3	0.671	psychologytoday.com
4	0.572	mayoclinic.org
5	0.418	healthline.com
6	0.390	e-monsite.com
7	0.317	cancer.org
8	0.305	heart.org
9	0.290	aarp.org
10	0.264	medicalnewstoday.com

3.2.2 WatSMM-CAL. In this run, we used our assessment system on the **M** collection to manually find useful documents for each topic. We primarily focused on finding useful documents and disregarded the correctness and credibility of the information.

For each topic, we initialize a CAL model with the topic's query as the seed query. One assessor spent a maximum of 10 minutes per topic using both "Search" and "Discovery" components to find documents. On average, we found 32.22 useful documents per topic (min=5, max=121). We used all the judgments made during a topic's session to build a classifier to score all documents in the collection. To produce the run, we rank documents based on the CAL models' scores.

3.2.3 WatSMM-CALHC. This run reranks the top 50 results from WatSMM-CAL based on the harmonic centrality [7] score of the hostnames, which are obtained from CommonCrawl [10]. Harmonic centrality originates from the social networks field and can be used to determine influencing nodes in a graph. Our goal for this run is to utilize harmonic centrality scores to push the most influential documents to the top of the list. We filtered the list of hosts only to contain the hostnames found in the **M** collection and normalized the scores to be between 0 and 1. Table 4 shows the top 10 hosts based on the normalized score. The top 50 documents from WatSMM-CAL are reranked based on the following:

$$\text{CAL score} \times (1 + \text{normalized harmonic centrality score})$$

Table 6 shows an example of the ordering of documents before and after our reranking method.

3.2.4 WatSMM-CALPR. This run follows the same procedure as WatSMM-CALHC, except it uses PageRank scores of hostnames, which were also obtained from CommonCrawl. Table 8 shows the top 10 hosts based on the normalized score.

3.2.5 WatSMM-Fused. This run uses reciprocal rank fusion [4] on runs WatSMM-CAL, WatSMM-CALHC, and WatSMM-CALPR.

3.2.6 WatSMC-CAL. This run is similar to WatSMM-CAL, except it is using the **C** collection that contains more documents. For each topic, we initialize a CAL model with the judgments from WatSMM-CAL as seed judgments. We primed the model for the new collection with an additional round of judgments with a maximum of 5 minutes

[10] https://commoncrawl.org/2019/11/host-and-domain-level-web-graphs-aug-sep-oct-2019/

Table 6. Example ordering based on harmonic centrality in WatSMM-CALHC for Topic #105 (Should I apply ice to burn?). While all results appear to contain useful information, the documents after the reordering allow more well-known websites (e.g. Mayo Clinic, WebMD, Healthline) to be ranked higher in the list.

Table 7. Before.

Rank	URL
1	http://eclinicalworks.adam.com/con...
2	https://melbournehandsurgery.com/o...
3	https://melbournehandsurgery.com/o...
4	https://melbournehandsurgery.com/h...
5	https://www.melbournehandsurgery.c...
6	http://rossa.kidshealth.org/en/par...
7	http://m.rossa-editorial.kidshealt...
8	http://m.rossa-editorial.kidshealt...
9	https://newsnetwork.mayoclinic.org...
10	https://www.healthychildren.org/En...

Table 8. After.

Rank	Before	URL
1	9	https://newsnetwork.mayoclinic.org...
2	16	https://www.webmd.com/first-aid/th...
3	15	https://www.healthline.com/health/...
4	6	http://rossa.kidshealth.org/en/par...
5	7	http://m.rossa-editorial.kidshealt...
6	8	http://m.rossa-editorial.kidshealt...
7	1	http://eclinicalworks.adam.com/con...
8	10	https://www.healthychildren.org/En...
9	11	https://www.healthychildren.org/en...
10	21	https://www.healthline.com/health/...

per topic. Topics were split between two assessors. Like WatSMM-CAL, we only focus on finding useful documents. On average, we found 60 useful documents per topic (min=14, max=236). Documents are ranked in a similar fashion as WatSMM-CAL.

3.2.7 WatSMM-CALQA100. For the six following runs, we experimented with using the RoBERTa language model [6] to rerank the results we obtained from our CAL models. RoBERTa, as with many transformer-based models, enforces a hard limit of 512 tokens as input. As such, using the whole document is often unfeasible. To work around this limitation, we select the most likely-relevant paragraph excerpt of the document to be used as part of the input to the language model. As shown in Table 9, these excerpts are often much shorter than the cap.

To find the paragraphs excerpts, we first split each document in the collection into a set of paragraphs using newlines as a delimiter. Excerpts are constructed such that they contain a minimum of 100 words while ignoring lines with five or fewer words to avoid including boilerplate content. For each topic, we created a new CAL model using our previous judgments as seeds to train the model, and instead of scoring documents like in the previous runs, we scored all generated paragraph excerpts and selected the topmost scoring paragraph for each document. Examples of these paragraphs are shown in Table 9. We use the topmost likely relevant paragraph excerpts to construct the top 1000 documents, with each document now having an associated most-likely relevant excerpt.

For our language model, we used `roberta-large` [6], and fine-tune it on the BoolQ (for Boolean Questions) dataset [3]. The BoolQ dataset contains natural language questions in the form of yes/no, with each question paired with a paragraph from Wikipedia containing the answer. We choose this dataset as it aligns with the track's goal of finding the correct information for different topics that are already written in the form of a yes/no question in the description field (e.g., "Does duct tape work for wart removal?"). Our goal was to use the topics' description field with the paragraph excerpts to determine if the answer matches with the stance field of the topic, i.e. the document provides the correct information. If a document provides the correct information, we change its position such that it is placed higher than documents with incorrect answers while still maintaining the original order. For this particular run, we only reranked the top 100 scoring documents from the CAL model trained on the **M** collection.

We used a batch size of 8, a learning rate of 1e-5, and three training epochs for fine-tuning the model. A softmax layer is applied to get the yes/no answer probabilities. We assigned the final answer as "yes" if the probability

Table 9. Example paragraph excerpts for few topics.

Topic: 104	Topic description: **Does duct tape work for wart removal?** (Stance: unhelpful) Q: Does duct tape work on common warts? A: Occasionally recommended as a home remedy for warts, duct tape has not been confirmed as an effective treatment. Research is conflicting, but some people believe that doing the following may help to get rid of a common wart: Covering the wart with a small piece of duct tape Removing the duct tape every three to six days and gently using an emery board or pumice stone on the wart Covering the wart with a fresh piece of duct tape about 10 to 12 hours later Results may only be seen after a number of weeks, if at all. Duct tape can cause skin irritation, bleeding and pain when removed. It should never be used in sensitive areas, such as the underarms or face.[28] *(docno: en.noclean.c4-train.07150-of-07168.4483. See more of this document in Figure 1b)*
Topic: 105	Topic description: **Should I apply ice to a burn?** (Stance: unhelpful) With a burn caused by a chemical, make sure the chemical or any clothing or jewelry in contact with the chemical is removed. If possible, use gloves so that you don't get burned elsewhere or so that someone helping you doesn't get burned. Put the burn under cool running water long enough to reduce pain, which may take about 10 to 15 minutes. If running water isn't available, you can immerse the burn in cool water or apply a cool, wet compress. Don't put ice directly on the burn. Dry the area with a clean cloth and apply a sterile, lightly wrapped bandage. Don't apply ointments or butter to a burn, as these can hold heat in the skin causing further damage in addition to increasing the risk of infection. *(docno: en.noclean.c4-train.03543-of-07168.43352)*
Topic: 106	Topic description: **Can vitamin b12 and sun exposure together help treat vitiligo?** (Stance: helpful) Most people with vitiligo generally use vitamins and supplements in combination with other treatments. Some studies have shown that folic acid, B12, and sun exposure, when used together, can aid in re-pigmenting the skin. Consult your doctor for the appropriate dosages. A few supplement combinations can be dangerous when combined or when taken out of balance with one another. Common vitamin deficiencies in people with vitiligo include folic acid, B12, copper, and zinc. As a result, doctors may prescribe vitamin supplements to boost your immune system. Vitamin B12 with Folic Acid Studies focusing on vitamin B12 deficiencies and vitiligo show a high incidence of vitiligo among individuals with pernicious anemia, a condition that hinders B12 absorption. Nevertheless, no recent studies indicate that supplementing vitamin B12, or B12 with folic acid, will help skin pigmentation. *(docno: en.noclean.c4-train.04820-of-07168.113071)*

of a yes answer is > 0.5, and as "no" otherwise. We matched the final answer with the topic's stance field to determine whether or not an answer was correct.

3.2.8 `WatSMM-CALQAAll`. This run is similar to `WatSMM-CALQA100`, except we do not enforce a reranking cutoff and instead rerank the entire list of documents. While this approach may introduce more irrelevant documents to be ranked higher in the list, it should effectively lower the rank of incorrect documents that can potentially be harmful and should not be shown to the user.

3.2.9 `WatSMC-CALQA100`. This run is similar to `WatSMM-CALQA100`. The main difference is that we are using the **C** collection, and for training the CAL models, we used the two rounds of judgments from `WatSMM-CAL` and `WatSMC-CAL`.

3.2.10 `WatSMC-CALQAAll`. This run is similar to `WatSMC-CALQA100`, except, like `WatSMM-CALQAAll`, we do not enforce a reranking cutoff.

3.2.11 `WatSMC-CALQAHC1`. This run sorts the predicted correct documents in `WatSMC-CALQAAll` based on their harmonic centrality score. Other documents are kept in their original order. The goal of this run is to introduce more correct and credible documents to be at the top of the page.

3.2.12 `WatSMC-CALQAHC2`. This run is similar to the previous run, except we only modify the ranking of predicted correct documents with domains part of the **M** collection. In other words, we ignore documents that were added outside of the **M** collection and only focus on sorting those that are part of the **M** collection. The intuition behind this method is that documents from the **M** collection are HONCode certified websites, and could be more reliable than the other websites in the **C** collection that were added using BM25.

3.2.13 `WatSMT-SD-S1`. The intuition behind the next two runs (`WatSMT-SD-S1` and `WatSMT-SD-S2`) is that misinformation carries different stance from the "truth". Therefore, given the correct stance towards a topic, we can train a model to detect the stance of a document and compare it with the correct stance to determine whether this document is misinformation or not.

We obtained a binary classification model by fine-tuning T5-Large [10] on a loosely balanced subset of effectiveness judgments from 2019 qrels (around 400 training examples). To exploit this text-to-text transformer, we constructed the input in a way similar to the approach by [8]: "stance detection topic: " + query + "document: " + document content.

Due to the 512 input tokens limit of T5, we summarized the document by selecting sentences that were most relevant to the topic. We scored each sentence based on a list of stance words[11] and the query terms. Specifically, for each word in the sentence, we first stemmed it and checked if it was among the stemmed stance words or the stemmed query terms. Each stance word would have a score of 1. Each query term would have a score relative to their position in the query because we wanted sentences to be more relevant to the treatment instead of the health issue. For example, for the query "yoga asthma", "yoga" would have a score of 2 and "asthma" would have a score of 1. However, in our post-TREC analysis, this query scoring scheme does not yield significantly better model performance than all query term with the equal score of 1. Then for each document, we selected those top-ranked sentences and concatenated them together in their original order in the document to form the input sequence of 512 words. We fine-tuned the model using the AdamW optimizer with a learning rate of 2e-5 and batch size of 16, with Early Stopping based on the F1-macro on the validation set (10% of the training set) with a patience of 3.

Similar to the work from [9], we applied a Softmax function on the logits of the word "favor" and the word "against" at the first generated token to get binary classification probabilities. We further mapped them into a single correct_probability by incorporating the "correct stance" field of each topic. That is, if the given topic was helpful, we took the favor_probability as the correct_probability. Otherwise, we used the against_probability. We applied the stance model trained above to predict the stance of the top 3,000 documents of each topic from this year's `baselineBM25` and reranked those documents using two strategies.

[11] Stance words: help, treat, benefit, effective, safe, evidence, improve, harm, hurt, useful, prove, ineffective, limit, poor, lack, insufficient, consider, quality, against, reliable.

WatSMT-SD-S1 and WatSMT-SD-S2 differ in the strategy to combine the BM25_score and the correct_probability. For this run, we used the following strategy to promote correct documents and suppress incorrect documents:

$$\text{final\_score} = \text{BM25\_score} \times e^{\text{correct\_probability}-0.5}$$

3.2.14 WatSMT-SD-S2. This run used a different reranking strategy from the one used in WatSMT-SD-S1. We preferred to order documents as: (1) documents with a clear and correct stance, (2) documents without a clear stance, and (3) documents with a clear but incorrect stance. Specifically, we used the following rule:

$$\text{final\_score} = \begin{cases} \text{BM25\_score} \times 10, & \text{if correct\_probability} > 0.75 \\ -\text{BM25\_score}, & \text{if correct\_probability} < 0.25 \\ \text{BM25\_score}, & \text{otherwise} \end{cases}$$

3.2.15 WatSMC-Correct. In this run, we manually assessed documents for usefulness and correctness using our high-recall retrieval system's "Search" and "Discovery" components [2]. The idea behind the WatSMC-Correct run was that training CAL on correct documents would allow us to find other correct documents while avoiding incorrect documents. Having the CAL model be trained on only correct documents would allow it to more easily learn the problem of finding correct documents than to let it learn incorrect documents as well, since the difference between correct and incorrect documents is mostly only the stance which is only a few words. For this run, we trained CAL on useful and correct documents, unlike the other CAL-based runs which ignored stance. Thus, we can compare performances between our different CAL-based approaches. At the same time, we can use our WatSMC-Correct run to better evaluate our other automatic and manual runs.

For the first judgement round, two assessors were restricted to using the "Search" interface (interactive search and judging) for approximately 10 minutes per topic to judge documents. Here, our definition of usefulness was different from our previous runs. A document was judged as "highly useful" if it contained an answer to the health issue that matched the stance given by the track for the topic, and assessors could mark other documents as "not useful" to keep track that they had viewed the document and decided it either was not useful or it was useful but incorrect. For the second round of judgement, the two assessors used the "Discovery" interface (CAL), with the first round's judgements for each topic used as seeds for training. The two assessors spent approximately 10 minutes each per topic to judge documents in this round. To speed up the assessment phase, the assessors were presented with most likely paragraph excerpt of the document to make their judgements. A document was marked as "highly useful" if its summary contained a correct answer. If the summary represented the document as one that was useful but contained an incorrect answer or lacking an answer, it was marked as "useful". Lastly, if the summary was not useful, then the document was marked as "not useful".

Figure 2 shows the total number of correct documents found by assessors for each topic which was used to train the final CAL model and were placed at the top of the run. The figure also shows that for each topic, the total number of correct documents is comprised of documents found using "Search" (interactive search and judging), denoted by the green bar, and additional correct documents found using "Discovery" (CAL), denoted by the blue bar. In total, 1481 correct documents were found by assessors across all topics, with 429 correct documents found using "Search" and 1052 additional correct documents found using "Discovery".

Both assessors agreed that it was harder to find correct documents for topics with an "unhelpful" stance compared to the topics with a "helpful" stance. Additionally, while using "Discovery" (CAL), many near-duplicate documents were returned to the assessors for judgement.

To create the run, the documents judged highly useful in both rounds of judging were placed first, followed by documents returned by the final CAL model trained on "highly useful" documents only (the other judgements were ignored for training the CAL model).

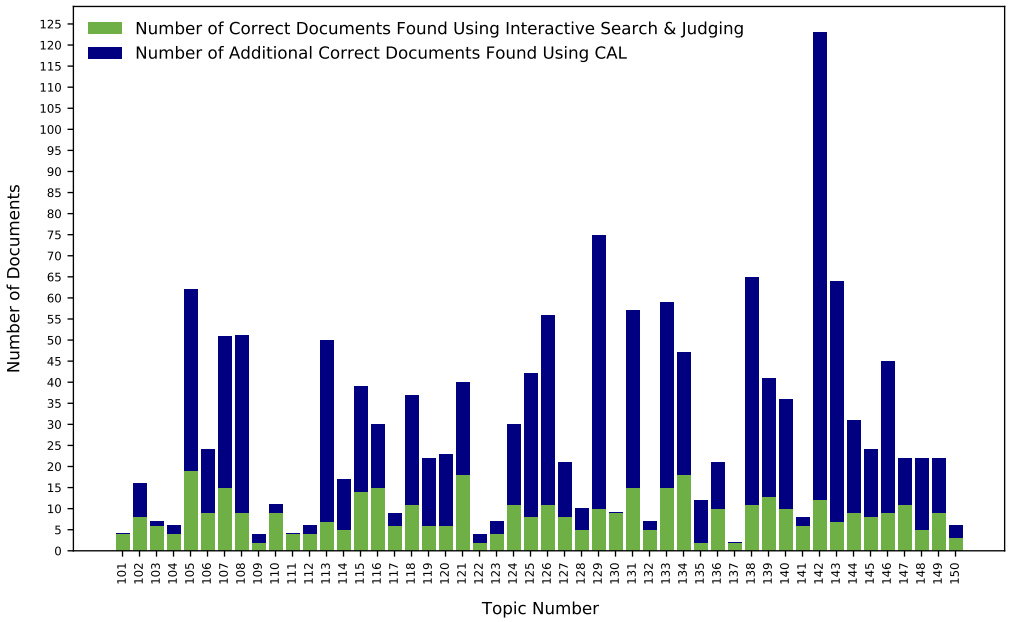


Fig. 2. This plot shows the total number of correct documents found by our assessors for each topic for the `WatSMC-Correct` run. Additionally, for each topic, the green bar indicates the number of correct documents found using "Search" (interactive search and judging) and the blue bar indicates the number of additional correct documents found using "Discovery" (CAL).

4 RESULT AND DISCUSSION

Coverage refers to the percentage of documents that have been assessed. Table 10 shows the coverage of assessed documents among top k documents in each run. We can argue that metrics focusing on the top 20 and fewer documents should be fair and objective among our runs, where we don't need to consider the case of useful/credible/correct documents not being assessed.

4.1 Automatic Runs

Table 11 shows our automatic BM25 run on the **M** collection (WatSAM-BM25) and the **E** collection (WatSAE-BM25) performing better than the baseline in terms of compatibility (0.055 and 0.042 vs. -0.015). In Figure 3, we can see the increase in helpful and decrease in harmful results for the filtered collections. It appears that using a refined dataset for retrieval can be an effective approach in reducing harmfulness and increasing helpfulness. It also appears that expanding **M** using the common crawl domain link graph and filtering out non-medical web pages to construct collection **E** further improved compatibility over the baseline with collection **E** giving a boost to helpful at the cost of a slight increase in harmful results.

In Table 11, we also see a significant jump for precision at rank 10 in WatSAE-BM25 and WatSAM-BM25 for credible documents compared to the baseline (0.557 and 0.586 vs. 0.417). This increase shows that collection filtering can help us retrieve credible documents. These runs also have an increase incorrect results as well (0.288 and 0.288 vs. 0.203). We also see a drop in the number of incorrect documents retrieved by the two

methods compared to the baseline (0.194 and 0.209 vs. 0.291). It appears that in addition to returning more correct information, the collections help in reducing the amount of incorrect information returned to the users.

The goal of creating **E** was to create a smaller, more reliable and more encompassing collection than **M**, which could aid in downstream tasks such as acting as a source of truth. While results show that the methods used were effective, further refinements need to be made before using this collection to detect the stance of a query automatically.

Unfortunately, further modifications to the WatSAE-BM25 run worsened results. Relevance feedback did not improve compatibility (0.031 vs. 0.055). Our reranking also worsened results (0.027 vs. 0.055). This is likely due to limited tuning data. As discussed in 2.2.3, previous years did not have a domain link graph, so using them as training data was not as effective as they were missing certain features.

4.2 Credibility-based Filtering and Reranking

Our goal for the **M** collection was to create a subset of the collection that only contains documents from reliable sources. In this collection, we used HONCode certification as our method of determining reliability, in addition to manually adding a few reliable websites. The certification assessment is done by medical experts that determined such websites to be reliable in providing medical information. As such, in our first five manual runs, we primarily focused on returning documents that were deemed useful and anticipated their correctness in providing health information to match the truth.

Table 13 shows precision scores under different criteria. Our baseline run, `baselineBM25` differs from `WatSAM-BM25` only in the type of collection used. In `WatSAM-BM25`, we used the **M** collection that contains reliable websites. In terms of finding correct documents, our baseline run and `WatSAM-BM25` seems to have about the same number of correct documents in the top 10 results, but `WatSAM-BM25` performs better in terms of returning credible documents in terms of precision@10. This difference in credibility is statistically significant using a two-tailed paired t-test ($p = 0.008$).

In terms of the overall compatibility measure (Table 11), all of our automatic runs perform better than the baseline, with our `WatSAE-BM25` run, which uses our smallest collection **E**, performing the best. Overall, the results indicate that using traditional retrieval methods with the filtering techniques described in Section 2.2 provides better performance than simply using the entire collection.

In `WatSMM-CAL`, we used our high-recall system to retrieve as many useful documents as possible from the **M** collection.

We also experimented with reranking based on harmonic centrality and PageRank. Both `WatSMM-CALHC` and `WatSMM-CALPR` attempt to push more influential to the top of the list. There is a slight increase in credibility over `WatSMM-CAL`, but the result is not statistically significant. All runs seem to have similar scores in terms of overall compatibility measure (Table 11), with `WatSMM-CALPR` having slightly lower performance.

4.3 Correctness and Stance detection-based reranking

We also experimented with reranking the results from the CAL models using the RoBERTa language model. The language model was fine-tuned on the BoolQ dataset, which contains questions in the form of yes/no. The goal of using the language model is to determine whether a relevant excerpt from a document contains an answer that matches with the topic's stance (i.e., whether the document has correct or incorrect information). Both `WatSMM-CALQA100` and `WatSMM-CALQAAll` attempts to rerank the results based on correctness under the **M** collection. In terms of helpful compatibility, both appear to perform the same, but reranking the complete set of results, as in `WatSMM-CALQAAll`, appears to lower the harmful compatibility score. The results are also similar for `WatSMC-CALQA100` and `WatSMC-CALQAAll` with the **C** collection. Under this reranking method, our best run in terms of helpful compatibility is `WatSMC-CALQAHC1`, where we further rerank the documents marked as correct

based on harmonic centrality, in an attempt to push more influential and correct documents to the top of the list. In terms of precision@10, this run has the highest score for documents that are useful & correct & credible. Compared to runs that used CAL scores alone, the runs that used the RoBERTa-based reranking method have better overall compatibility performance.

For `WatSMT-SD-S1` and `WatSMT-SD-S2`, we applied a stance detection model to promote correct documents and suppress incorrect documents. Among our manual runs, those two are the only ones that don't utilize manual assessments, although they need the correct stance toward each health treatment. From the performance in Tables 11 and 13, we can see that `WatSMT-SD-S1` and `WatSMT-SD-S2` are comparable with those other manual runs in terms of correctness. Specifically, if we focus on the Compatibility measure (helpful - harmful) in Table 11, we can notice that `WatSMT-SD-S1` is the most competitive run approaching human level (0.183 v.s. 0.226). This fact also demonstrates the power of pre-trained language models applied to stance detection tasks. In the future, there is still much room for further improving the classification performance of the stance detection model. Meanwhile, we can also add a reranking stage between the BM25 stage and the stance detection stage to improve the relevance and credibility of top results.

4.4 `WatSMC-Correct` Run

The `WatSMC-Correct` run was created by having assessors manually judge documents to include only correct documents into the training set for CAL. As such, we expected that the run would perform well. Results show that the run did indeed perform generally well across all metrics. It was our best run as per the Compatibility (helpful - harmful), Compatibility (helpful), nDCG (Useful & Correct), nDCG (Useful & Correct & Credible) and P@10 (Useful & Correct) measures. However, under some other metrics, there are other runs that performed better than the `WatSMC-Correct` run. This could be because assessors only had a limited time to make assessments and/or for some topics, it was hard for assessors to manually find many correct documents. Another observation we found from the results was that even though assessors were able to find at least two correct documents for every topic, the P@10 score for a few topics was still 0. Lastly, as the motivation of this run was to help us better evaluate our other runs, we can see that some of the other runs did a comparable job to what humans can do.

5 CONCLUSION

In this report, we introduced several methods to tackle the challenge of misinformation in online health searches.

To retrieve correct and credible results, we constructed curated collections based on the URL domain credibility. Running BM25 on these collections achieved higher compatibility scores compared to the baseline, returning more helpful and fewer harmful documents.

We also used continuous active learning to find useful documents within these curated collections, relying on the idea that these collections should contain credible information. This approach resulted in a higher compatibility score than the baseline and BM25 runs on those same curated collections.

Our runs utilizing the correct topic stance showed the value of using pre-trained language models to detect the stance of documents. Runs using this approach showed a sizeable drop in the number of harmful documents returned as well as an increase in the number of helpful documents returned.

ACKNOWLEDGMENTS

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (RGPIN-04665-2020, RGPAS-00080-2020), in part by Google, in part by Compute Canada, and in part by the University of Waterloo.

Table 10. Assessment coverage of the runs.

	Run tag	Top 10	Top 20	Top 50	Top 1000
	baselineBM25	1.0	1.0	0.841	0.189
Automatic	WatSAM-BM25	1.0	1.0	0.650	0.070
	WatSAE-BM25	1.0	1.0	0.718	0.078
	WatSAE-BM25RM3	1.0	1.0	0.627	0.074
	WatSAE-BM25-RR	1.0	1.0	0.687	0.078
Manual	WatSMM-CAL	1.0	1.0	0.668	0.077
	WatSMM-CALHC	1.0	1.0	0.668	0.077
	WatSMM-CALPR	1.0	1.0	0.668	0.077
	WatSMM-Fused	1.0	1.0	0.668	0.077
	WatSMC-CAL	1.0	1.0	0.789	0.159
	WatSMM-CALQA100	1.0	1.0	0.669	0.077
	WatSMM-CALQAAll	1.0	1.0	0.628	0.077
	WatSMC-CALQA100	1.0	1.0	0.715	0.170
	WatSMC-CALQAAll	1.0	1.0	0.735	0.170
	WatSMC-CALQAHC1	1.0	1.0	0.739	0.170
	WatSMC-CALQAHC2	1.0	1.0	0.737	0.170
	WatSMT-SD-S1	1.0	1.0	0.693	0.129
	WatSMT-SD-S2	1.0	1.0	0.722	0.136
	WatSMC-Correct	1.0	1.0	0.659	0.141

Table 11. Performance of the runs using the compatibility measure (* indicates statistical significance computed over the baselineBM25 run at $p < 0.05$).

	Run tag	Compatibility (helpful)	Compatibility (harmful)	Compatibility (helpful - harmful)
	baselineBM25	0.129	0.144	-0.015
Automatic	WatSAM-BM25	0.161	0.119	0.042
	WatSAE-BM25	**0.178***	0.123	**0.055**
	WatSAE-BM25RM3	0.140	**0.109**	0.031
	WatSAE-BM25-RR	0.145	0.118	0.027
Manual	WatSMM-CAL	0.194*	0.129	0.065
	WatSMM-CALHC	0.201*	0.136	0.065
	WatSMM-CALPR	0.197*	0.139	0.058
	WatSMM-Fused	0.203*	0.140	0.063
	WatSMC-CAL	0.225*	0.167	0.058
	WatSMM-CALQA100	0.208*	0.064	0.144
	WatSMM-CALQAAll	0.203*	**0.034***	0.169
	WatSMC-CALQA100	0.217*	0.134	0.083
	WatSMC-CALQAAll	0.234*	0.080	0.154
	WatSMC-CALQAHC1	0.248*	0.079*	0.169
	WatSMC-CALQAHC2	0.225*	0.055*	0.170
	WatSMT-SD-S1	0.220*	0.037*	0.183
	WatSMT-SD-S2	0.196*	0.059*	0.137
	WatSMC-Correct	**0.281***	0.055*	**0.226**

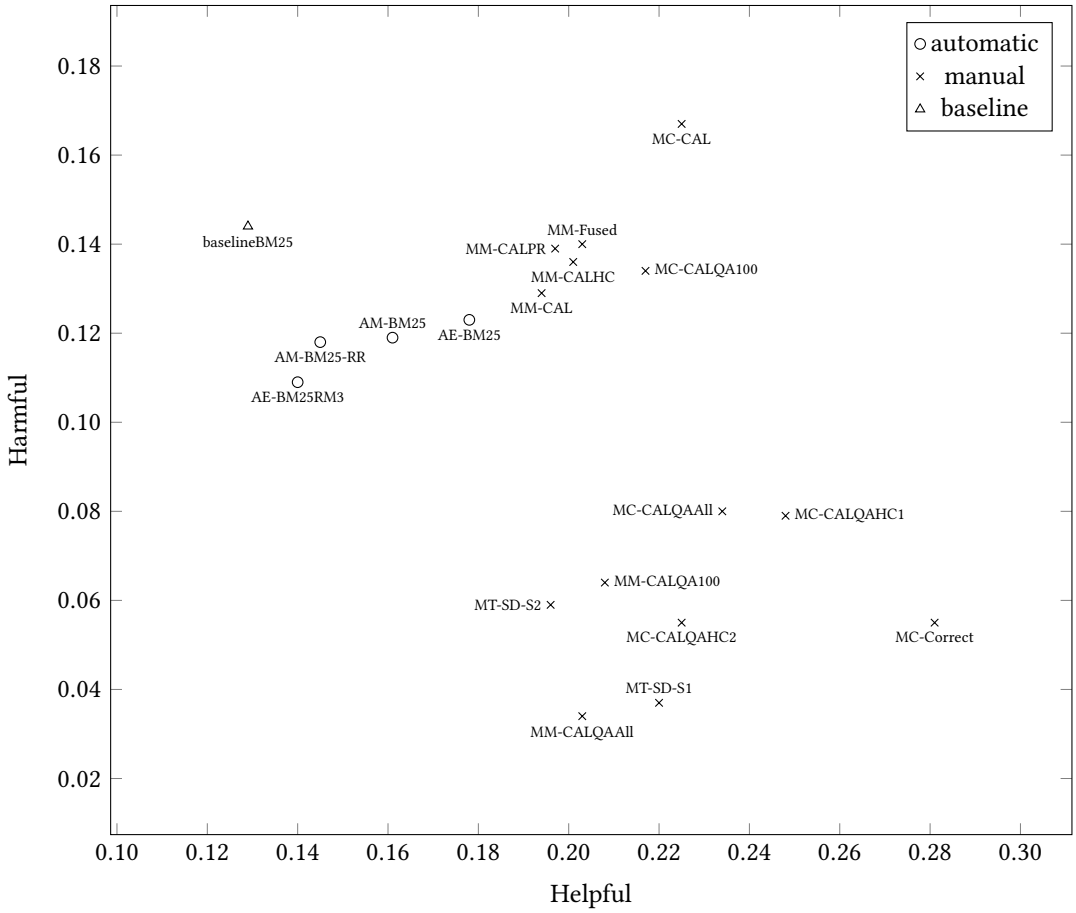


Fig. 3. Helpful-Harmful Compatibility plot. "WatS" has been dropped from run names. Runs in the lower right corner have a better compatibility metric. The cluster of runs in the lower right all make use of the topic stance. Note the dramatic shift in compatibility between WatSMC-CAL and WatSMC-Correct. This shows that training for correctness can greatly improve compatibility.

REFERENCES

[1] Mustafa Abualsaud, F. C. Beylunioğlu, M. Smucker, and P. R. Duimering. 2019. UWaterlooMDS at the TREC 2019 Decision Track. In *TREC*.

[2] Mustafa Abualsaud, Nimesh Ghelani, Haotian Zhang, Mark D. Smucker, Gordon V. Cormack, and Maura R. Grossman. 2018. *A System for Efficient High-Recall Retrieval.* Association for Computing Machinery, New York, NY, USA, 1317–1320. https://doi.org/10.1145/3209978.3210176

[3] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. arXiv:cs.CL/1905.10044

[4] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09).* Association for Computing Machinery, New York, NY, USA, 758–759. https://doi.org/10.1145/1571941.1572114

[5] Taher H Haveliwala. 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering* 15, 4 (2003), 784–796.

Table 12. Performance of the runs using the nDCG measure. An asterisk indicates statistical significance computed over the baselineBM25 run at $p < 0.05$ using a two tailed paired t-test. For each metric we highlight the best result for automatic and manual runs.

	Run tag	Useful & Correct	Useful & Credible	Useful & Correct & Credible	Incorrect
	baselineBM25	0.428	0.487	0.381	0.381
Automatic	WatSAM-BM25	0.164*	0.259*	0.199*	0.134*
	WatSAE-BM25	**0.189***	**0.296***	**0.226***	0.147*
	WatSAE-BM25RM3	0.170*	0.272*	0.196*	**0.133***
	WatSAE-BM25-RR	0.183*	0.284*	0.219*	0.143*
Manual	WatSMM-CAL	0.179*	0.284*	0.212*	0.129*
	WatSMM-CALHC	0.179*	0.289*	0.212*	0.129*
	WatSMM-CALPR	0.174*	0.284*	0.208*	0.134*
	WatSMM-Fused	0.178*	0.290*	0.213*	0.131*
	WatSMC-CAL	0.448	0.546	0.411	0.362
	WatSMM-CALQA100	0.200*	0.279*	0.241*	0.109*
	WatSMM-CALQAAll	0.199*	0.265*	0.241*	**0.092***
	WatSMC-CALQA100	0.488	**0.547**	0.450	0.341
	WatSMC-CALQAAll	0.505*	0.537	0.462*	0.311*
	WatSMC-CALQAHC1	0.495	0.542	0.463*	0.310*
	WatSMC-CALQAHC2	0.470	0.533	0.447	0.299*
	WatSMT-SD-S1	0.495*	0.372*	0.413*	0.112*
	WatSMT-SD-S2	0.467	0.377*	0.398*	0.114*
	WatSMC-Correct	**0.520***	0.498	**0.517***	0.184*

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:cs.CL/1907.11692

[7] Massimo Marchiori and Vito Latora. 2000. Harmony in the small-world. *Physica A: Statistical Mechanics and its Applications* 285, 3-4 (Oct 2000), 539–546. https://doi.org/10.1016/s0378-4371(00)00311-3

[8] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 708–718. https://doi.org/10.18653/v1/2020.findings-emnlp.63

[9] Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. *Vera: Prediction Techniques for Reducing Harmful Misinformation in Consumer Health Search.* Association for Computing Machinery, New York, NY, USA, 2066–2070. https://doi.org/10.1145/3404835.3463120

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[11] Haotian Zhang, Mustafa Abualsaud, Nimesh Ghelani, Angshuman Ghosh, Mark Smucker, Gordon Cormack, and Maura Grossman. 2017. UWaterlooMDS at the TREC 2017 Common Core Track. In *Text Retrieval Conference (TREC)*. https://trec.nist.gov/pubs/trec26/papers/UWaterlooMDS-CC.pdf

Table 13. Performance of the runs using precision measure (* indicates statistical significance computed over the baselineBM25 run at $p < 0.05$).

	Run tag	Useful		Useful & Correct		Useful & Credible		Useful Correct & Credible		Incorrect	
		P@10	P@50	P@10	P@50	P@10	P@50	P@10	P@50	P@10	P@50
	baselineBM25	0.694	0.569	0.309	0.259	0.417	0.307	0.203	0.142	0.291	0.212
Automatic	WatSAM-BM25	0.686	0.357*	0.303	0.142*	0.557*	0.270	0.288*	0.121	0.194	0.099*
	WatSAE-BM25	0.731	0.414*	0.309	0.174*	0.586*	0.310	0.288*	0.141	0.209	0.115*
	WatSAE-BM25RM3	0.646	0.369*	0.300	0.146*	0.477	0.270	0.248	0.116	0.172*	0.109*
	WatSAE-BM25-RR	0.654	0.383*	0.309	0.156*	0.543*	0.286	0.288*	0.125	0.184	0.112*
Manual	WatSMM-CAL	0.743	0.454*	0.376	0.218	0.626*	0.373	0.339*	0.197	0.181*	0.116*
	WatSMM-CALHC	0.740	0.454*	0.374	0.218	0.654*	0.373	0.345*	0.197	0.191*	0.116*
	WatSMM-CALPR	0.771	0.454*	0.376	0.218	0.646*	0.373	0.339*	0.197	0.203	0.116*
	WatSMM-Fused	0.757	0.454*	0.382	0.218	0.643*	0.373	0.355*	0.197	0.203	0.116*
	WatSMC-CAL	0.897*	0.649*	0.506*	0.337*	0.620*	0.456*	0.364*	0.250*	0.250	0.188
	WatSMM-CALQA100	0.666	0.430*	0.397	0.209	0.551*	0.345	0.361*	0.183	0.091*	0.097*
	WatSMM-CALQAAll	0.600	0.358*	0.397	0.204	0.500	0.284	0.361*	0.176	0.041*	0.041*
	WatSMC-CALQA100	0.846*	0.606	0.497*	0.345*	0.597*	0.377*	0.379*	0.228*	0.237	0.168
	WatSMC-CALQAAll	0.786	0.588	0.515*	0.368*	0.557*	0.364	0.391*	0.243*	0.138*	0.116*
	WatSMC-CALQAHC1	0.780	0.584	0.518*	0.358*	0.566*	0.382	0.409*	0.255*	0.128*	0.116*
	WatSMC-CALQAHC2	0.669	0.517	0.406	0.272	0.526	0.354	0.361*	0.205	0.078*	0.116*
	WatSMT-SD-S1	0.711	0.469*	0.497*	0.345*	0.414	0.285	0.291*	0.206*	0.103*	0.060*
	WatSMT-SD-S2	0.674	0.467*	0.453*	0.321*	0.423	0.279	0.276*	0.187*	0.119*	0.073*
	WatSMC-Correct	0.826*	0.502	0.568*	0.334*	0.589*	0.342	0.403*	0.234*	0.091*	0.068*