Overview of the TREC 2021 Health Misinformation Track

Charles L. A. Clarke[1], Maria Maistro[2], and Mark D. Smucker[1]

[1]University of Waterloo
[2]University of Copenhagen

1 Introduction

TREC 2021 was the third year for the Health Misinformation track, which was named the Decision Track in 2019 [1]. In 2021, the track had an ad-hoc retrieval task. In each year, the track has used a crawl for its document collection. In 2019 and 2021, we used web crawls, and in 2020, we used a web crawl restricted to news sites.

By focusing on health-related ad-hoc web search, the track brings new challenges to the web retrieval task. The most striking difference is that for health search, documents containing incorrect information are considered to be harmful and not merely non-relevant. As such, retrieval systems need to actively work to avoid including or ranking this incorrect, harmful information highly in the results. For relevant documents that contain correct information, we prefer sources with higher credibility.

This year, each topic's description was expressed as a question, for example "Should I apply ice to a burn?". A topic also has a query, for example "put ice on a burn", that represents what a user might enter if they do not ask a full question. All topics concern themselves with determining the efficacy of a treatment for a health issue. Based on a credible source of information, we declare a *stance* for a topic as either *helpful* or *unhelpful*. We provide an *evidence* URL link to the source we used to determine the stance. Each topic is also supplied with a narrative providing additional clarification to the assessors.

Automatic runs could only make use of the topic's query or description. If a run used the narrative, stance, or evidence, it had to be considered a manual run. A challenge of health-related search is determining what is correct information, i.e., determining the correct stance for a topic.

Based on the assessors' judgments, we establish a preference ordering for documents considered to be helpful as well as for documents considered to be harmful. Helpful documents are supportive of helpful treatments or try to dissuade the reader from using unhelpful treatments. Harmful documents encourage use of unhelpful treatments or dissuade the reader from using helpful treatments. Whether a treatment is considered helpful or unhelpful is based on our provided stance.

Submitted runs are evaluated based on their *compatibility* [4, 5] with both a preference ordering for helpful documents as well as a preference ordering for harmful documents. The best runs have high compatibility with the helpful preference ordering and low compatibility with the harmful ordering. The preference orderings take into consideration the usefulness, correctness, and credibility of the documents.

2 Topics

We created 50 topics this year with half of them having a stance of helpful and half with a stance of unhelpful. Figure 1 shows an example of a topic.

NIST was only able to provide assessments for 35 of the 50 topics. Of these 35 topics, no harmful documents were found for topics 127, 133, and 145. We have excluded these three topics from the analysis in this paper.

```
<topic>
<number>105</number>
<query>put ice on a burn</query>
<description>Should I apply ice to a burn?</description>
<narrative>Many people commonly put on ice on burns in an attempt to
stop the burning and pain. A very useful document would discuss the
effectiveness of using ice to treat burns. A useful document would
help a user decide if putting ice on burns is a recommended treatment
by providing information on recommended treatments for burns and may
not discuss ice as a treatment, or a useful document may discuss
benefits or concerns for application of ice to skin.</narrative>
<disclaimer>We do not claim to be providing medical advice, and
medical decisions should never be made based on the stance we have
chosen. Consult a medical doctor for professional advice.</disclaimer>
<stance>unhelpful</stance>
<evidence>https://www.uwhealth.org/news/the-right-way-to-treat-burns</evidence>
</topic>
```

Figure 1: Example of a topic for the TREC 2021 Health Misinformation track.

3 Document Collection

This year we used the noclean version of the C4 dataset[1] used by Google to train their T5 model. The collection is comprised of plain text extracted from the April 2019 snapshot of the Common Crawl and contains over 1 billion English web pages. The noclean version of C4 was used rather than the clean version to provide the full text of a web page. We observed many cases where the clean version of C4 removes section headers and important material. The clean version of C4 is designed for training a language model, which is a different purpose than retrieval.

4 Submitted Runs

Seven groups submitted 71 runs to the adhoc retrieval task. The UWaterlooMDS group submitted a BM25 baseline run, `baselineBM25`, which used the topic's query field and default parameters from Anserini ($k_1 = 0.9$, $b = 0.4$, Porter stemming, stopword removal). Table 1 reports an overview of the participating groups and the number of runs submitted by each group. Next we present a brief summary of the approach adopted by each group.

[1] `https://huggingface.co/datasets/allenai/c4`

[2] Includes one baseline run.

Table 1: Overview of the groups participating in the TREC Health Misinformation track 2021.

Group Name	Organization	# Submitted Runs
CiTIUS	University of Santiago de Compostela	10
DigiLab	University of Geneva	7
h2oloo	University of Waterloo (Lin)	10
UPV	Valencia Polytechnic University	10
UWaterlooMDS[2]	University of Waterloo (Smucker)	19
Waterloo_Cormack	University of Waterloo (Cormack)	9
Webis	Bauhaus University, Weimar	6

CiTIUS [7] used BM25 as base ranker and different strategies to perform passage re-ranking of the top 100 documents. Passage re-ranking was performed with respect to the original topic or hand-crafted expressions generated from the topic's fields. RoBERTa was used to represent sentences and compute the similarity of the passages within the top 100 documents and the topic. An additional classifier for passage reliability was trained on data from past editions. Finally, scores from different components were merged with CombSUM or Borda Count.

DigiLab [12] implemented a two-step ranking approach that includes a standard retrieval phase, based on the BM25 model, and a re-ranking phase, with a pipeline of models to estimate (1) usefulness, (2) supportiveness, and (3) credibility. The usefulness ranking was generated with a set of transformer-based language models fine-tuned on the MS MARCO corpus. The supportiveness ranking was generated with BERT-based models fine-tuned on scientific and Wikipedia corpora. The credibility ranking was generated with a random forest model trained on the Microsoft Credibility dataset combined with a list of credible sites. The resulting ranked lists were fused with Reciprocal Rank Fusion (RRF).

h2oloo used Pyserini's default BM25 as base ranker. Re-ranking was performed with a combination of different T5 models (mono and duo) and Vera [10] with different topic fields.

UPV [11] also used Pyserini's default BM25 as base ranker. Usefulness scores were estimated as the similarity between documents represented with Bio Sentence BERT and the topic's description. Credibility was estimated with cosine similarity between documents represented with RoBERTa and a reference document satisfying 4 different credibility criteria. Finally, BM25 scores, usefulness scores, and credibility scores were fused in a single ranking.

UWaterlooMDS [2] submitted 19 runs (5 automatic and 14 manual). One of their runs, baselineBM25 was used by the track organizers as a baseline for the task. Their automatic runs focused on experimenting with creating subcollections of higher quality and then performing retrieval over these subcollections. Their manual runs used 3 different approaches. The first manual approach applied continuous active learning (CAL) to find relevant documents over one of their filtered subcollections. The second approach reranked the output of CAL with RoBERTa, fine tuned on BoolQ dataset, to detect the stance of the document. The final approach used BM25 as the base ranker and then used the T5-large model to re-rank the top 3K results. T5 was fine-tuned on a balanced subset of 2019 qrels to predict the stance of each document, and different strategies were used to fuse BM25 and T5 scores.

Waterloo_Cormack trained a logistic classifier with search results returned by Google or medline BM25. In some cases the term "Pubmed" was added to the topic's query as additional search term. Reciprocal rank fusion (RRF) was used to fuse different combinations of the above.

Webis [3] exploited Anserini's BM25 and PyGaggle's default MonoT5 model to create two base-

line rankings. Then the top 20 documents of each baseline ranked list were re-ranked according to 3 argumentative axioms with different weighting schemes for queries that seem to be argumentative.

5 Evaluation

Runs were evaluated by using a script[3] to compute the compatibility measure [4, 5]. We derive a qrels file to use with compatibility from the original NIST qrels file.

5.1 qrels (query-relevance files)

NIST used the track's relevance assessing guidelines[4] to generate the track's qrels. The format adopted for NIST qrels file is as follows:

```
topic_id 0 doc_id usefulness-judgment supportiveness-judgment credibility-judgment
```

where the columns are space separated. Documents were assessed by NIST assessors with respect to 3 criteria, which were recorded in the NIST qrels as follows:

- *Usefulness*: does the document contain material that the search user might find useful in answering the topic's question? Usefulness was assessed on 3-point scale: 0 if the document is not useful, 1 if the document is useful, and 2 if the document is very useful. This is column 4 of the qrels file.
- *Supportiveness*: does the document contain information that supports/dissuades the use of the treatment in the question? Supportiveness could be assessed to be one of three values: 0 means that the document dissuades the use of the treatment, 1 means that the document neither dissuades or supports the use of the treatment (neutral), and 2 means that the document supports the use of the treatment. This is column 5 of the qrels file.
- *Credibility*: how credible is the document? Credibility was assessed on a on 3-point scale: 0 if the document has low credibility, 1 if the document has good credibility, but does not exhibit the highest quality and credibility, and 2 if the document is excellent, i.e., it exhibits the highest quality and most credible information source. This is column 6 of the qrels file.

Notes:

- The assessors were not to refer to the topic's stance while judging, and usefulness judgements do not depend on the credibility of the source.
- Credibility is judged based on the assessor's expert opinion. A set of guidelines was developed to guide assessors in judging credibility (reported in the assessing guidelines).
- When a document was judged as not useful, it was not judged for its supportiveness nor for its credibility (value -1). In some cases, a useful document was accidentally not judged for its answer or credibility, i.e., a "skip" (value -2).

[3] https://github.com/trec-health-misinfo/Compatibility

[4] https://trec-health-misinfo.github.io/docs/TREC-2021-Health-Misinformation-Track-Assessing-Guidelines_Version-2.pdf

- Even by reducing the pool depth to 20, NIST assessors were only able to judge 35 out of the 50 topics due to lack of time. The missing topics are: 113, 116, 119, 123, 124, 125, 126, 130, 135, 138, 141, 142, 147, 148, 150. Possible reasons for the increase of time needed for judgement might be: 1) differently from previous editions [1, 6], credibility was judged with a 3-point scale instead of a binary label; 2) the documents in the C4 dataset are difficult to read as text extracts, and 3) usefulness was possibly too broadly defined this year. This issue needs to be further investigated.

5.2 Derived qrels

We took the NIST qrels and generated derived qrels for the various evaluation measures. We describe this next.

5.2.1 Preference Levels

For the compatibility measure, we converted the 3 aspects judged for documents (usefulness, supportiveness, and credibility) into a basic preference ordering as reported in Table 2. A document with a higher preference value is preferred to a document with a lower preference value. To define the preference order among tuples of labels we decided to favour credibility over usefulness (assuming the same correctness label). For example, in Table 2, one could consider to rank 11 above 10 or to swap them: both documents are correct, but one is more useful while the other is more credible. Since we favour credibility, the more credible document comes first.

Assessors judged useful and very useful documents for their supportiveness towards the health treatment. A document's supportiveness could be judged as *supports*, *neutral*, or *dissuades*. Assessors did not judge the supportiveness or the credibility of not-useful documents.

A document is correct if it is supportive of helpful treatments or dissuades unhelpful treatments. A document is incorrect if it dissuades from helpful treatments and is supportive of unhelpful treatments. Note that neutral documents are neither correct nor incorrect.

It is tempting to use the preference values as scores to compute Normalized Discounted Cumulated Gain (nDCG), but that ignores the incorrect information, which is critical to understanding the quality of results. In addition, we do not have a notion of *gain* here, which is a critical component of nDCG. We only know that we prefer certain documents to other documents. Of particular note, we prefer not-useful documents to incorrect documents.

We use the preference ordering to create a set of helpful and harmful preference qrels. We create helpful qrels by taking all preference values greater than zero. To create the harmful qrels, we use only the absolute value of the negative scores. Thus, the most harmful documents are those that are judged to be very useful or useful, incorrect, and have excellent credibility.

With helpful and harmful preference orderings, we can compute a run's *compatibility* with helpful and harmful documents. A run wants high compatibility with helpful documents and low compatibility with harmful documents.

5.2.2 Graded and Binary Relevance

We created a series of qrels files in the standard qrels format for graded and binary relevance effectiveness measures. These files can be used to compute Convex Aggregating Mean (CAM) [8] and Multidimensional Measure (MM) [9]. We recommend use of compatibility with helpful and harmful documents as the primary measures.

Table 2: Preference ordering for documents.

Preference Value	Usefulness	Correctness	Credibility
12	Very Useful	Correct	Excellent
11	Useful	Correct	Excellent
10	Very Useful	Correct	Good
9	Useful	Correct	Good
8	Very Useful	Correct	Low or Not Judged
7	Useful	Correct	Low or Not Judged
6	Very Useful	Neutral or Not Judged	Excellent
5	Useful	Neutral or Not Judged	Excellent
4	Very Useful	Neutral or Not Judged	Good
3	Useful	Neutral or Not Judged	Good
2	Very Useful	Neutral or Not Judged	Low or Not Judged
1	Useful	Neutral or Not Judged	Low or Not Judged
0	Not Useful	Not Judged	Not Judged
-1	Very Useful or Useful	Incorrect	Low or Not Judged
-2	Very Useful or Useful	Incorrect	Good
-3	Very Useful or Useful	Incorrect	Excellent

- Usefulness. Ignores answer correctness and document credibility. Obtained from NIST qrels by dropping supportiveness and credibility columns.
- Binary Usefulness. Same as the above, but usefulness is mapped to binary labels with a lenient mapping: if the document is useful or very useful, then it is mapped to 1; not useful documents are still mapped to 0.
- Useful and credible. Note that a document cannot be judged credible unless it is judged useful. A document is credible if only judged to have good or high credibility, otherwise it is not credible.
- Useful and correct. Note that a document cannot be judged correct unless it is judged useful.
- Useful and correct and credible.
- Incorrect. A document is incorrect if it is useful and is against the topic's given stance (a neutral document is not incorrect).

5.2.3 Multiple Aspect qrels

We created three aspect qrels as follows. The correctness column is mapped to 1, if the document's supportiveness aligns with the topic stance, and to 0 otherwise (no distinction for not judged or neutral). The credibility column is the same except that a -1 (not judged) is mapped to 0 (not credible). We also created two aspect qrels but only consider usefulness and one of the other two aspects.

5.3 Evaluation Measures

We evaluate runs by their *compatibility* with helpful and harmful results.

6 Results

Tables 3 and 4 report the results for automatic and manual runs. Figure 2 plots runs' compatibility with helpful and harmful results. For two runs with the same level of compatibility with helpful results, the run with the lower compatibility with harmful results is to be preferred. Thus the vera_mdt5_0.5 and vera_mt5_0.5 runs are notable for having high compatibility with helpful and low compatibility with harmful results.

The baselineBM25 run had a helpful compatibility of 0.122 and a harmful compatibility of 0.144. It seems that when no effort is made to prefer correct documents over incorrect documents, search results can be a mix of both, which can have negative consequences for people looking to search engines to help them make a decision about the efficacy of a health treatment.

The best scoring automatic runs (mt5, all_use_sup_cre, WatSAE-BM25) have helpful compatibility ranging from 0.137 to 0.195 and harmful compatibility from 0.095 to 0.153. In comparison to the BM25 baseline, the mt5 run boosted its compatibility with helpful results, but mt5 also increased its compatibility with harmful results. Both all_use_sup_cre and WatSAE-BM25 were able to boost helpful compatibility and reduce harmful compatibility in comparison to baselineBM25. The mt5 run is a T5 (MedMonoT5) reranking of BM25 results. The all_use_sup_cre run used a fusion of runs produced using separate models for usefulness, supportiveness, and credibility. The WatSAE-BM25 run is BM25 run over a curated collection that aims to contain credible health documents.

In comparison to the automatic runs, the best manual runs (vera_mdt5_0.5 and vera_mt5_0.5) have a helpful compatibility of around 0.298 and a low compatibility with harmful results of around 0.040. These runs are produced by first manually reformulating the description field to align with the topic's stance and then reranking BM25 results using T5 (mono-duo-T5 and mono-T5, respectively) along with a reranking based on the stance (Vera). Thus, we may be able to infer that the vera_mt5_0.5's performance gains over mt5 come from the ability to promote correct results and demote incorrect results.

The WatSMC-Correct run had human assessors interactively search and use a continuous active learning (CAL) tool for finding correct documents, which they placed at their top ranks. While this run did not consider credibility in its ranking of documents, it represents a reasonable standard of what humans can achieve in a limited time. As such, the vera_mdt5_0.5 and vera_mt5_0.5 runs are impressive in their ability to exceed the performance of reasonable human effort based on our current evaluation. Unknown is what human performance would be if a human generated ranking also ordered documents based on credibility in addition to correctness.

7 Acknowledgments

Thanks to Mustafa Abualsaud, Irene XiangYi Chen, Kamyar Ghajar, Linh Nhi Phan Minh, Amir Vakili Tahami, and Dake Zhang for their contributions to the running of the track.

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), in part by Google, in part by the facilities of Compute Canada, and in part by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 893667. Any opinions, findings and conclusions or recommendations

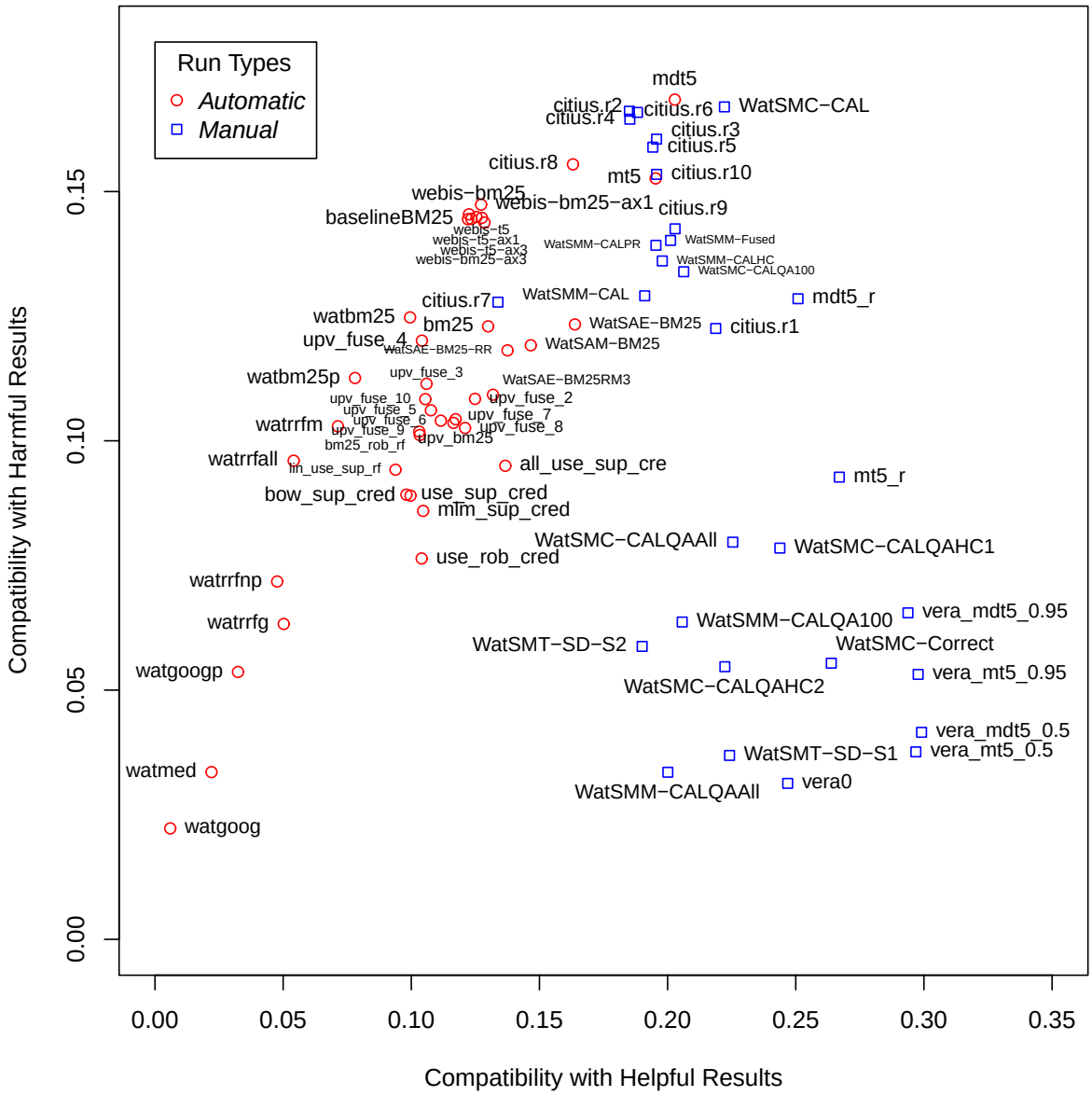

Figure 2: Compatibility of runs with helpful and harmful results. A good run is helpful and not harmful. For a given level of helpfulness, a run with less harm is to be preferred.

			Avg. Compatibility		
Group	Run	Topic Fields	help	harm	help-harm
h2oloo	mt5	desc	0.195	0.153	0.043
DigiLab	all_use_sup_cre	query, desc	0.137	0.095	0.042
UWaterlooMDS	WatSAE-BM25	query	0.164	0.123	0.040
h2oloo	mdt5	desc	0.203	0.168	0.034
DigiLab	use_rob_cred	query, desc	0.104	0.076	0.028
UWaterlooMDS	WatSAM-BM25	query	0.147	0.119	0.027
UWaterlooMDS	WatSAE-BM25RM3	query	0.132	0.109	0.023
UWaterlooMDS	WatSAE-BM25-RR	query	0.138	0.118	0.019
DigiLab	mlm_sup_cred	query, desc	0.105	0.086	0.019
UPV	upv_fuse_8	desc	0.121	0.103	0.018
UPV	upv_fuse_2	desc	0.125	0.108	0.016
UPV	upv_bm25	desc	0.117	0.104	0.013
UPV	upv_fuse_7	desc	0.116	0.104	0.013
DigiLab	bow_sup_cred	query, desc	0.100	0.089	0.011
DigiLab	use_sup_cred	query, desc	0.098	0.089	0.009
CiTIUS	citius.r8	query, desc	0.163	0.155	0.008
UPV	upv_fuse_6	desc	0.112	0.104	0.008
h2oloo	bm25	desc	0.130	0.123	0.007
DigiLab	bm25_rob_rf	query, desc	0.103	0.101	0.002
UPV	upv_fuse_5	desc	0.108	0.106	0.002
UPV	upv_fuse_9	desc	0.103	0.102	0.001
DigiLab	lin_use_sup_rf	query, desc	0.094	0.094	0.000
UPV	upv_fuse_10	desc	0.106	0.108	-0.003
UPV	upv_fuse_3	desc	0.106	0.111	-0.005
Waterloo_Cormack	watmed	query	0.022	0.034	-0.011
Waterloo_Cormack	watrrfg	query	0.050	0.063	-0.013
Webis	webis-t5-ax3	query	0.129	0.144	-0.015
UPV	upv_fuse_4	desc	0.104	0.120	-0.016
Waterloo_Cormack	watgoog	query	0.006	0.022	-0.016
Webis	webis-t5	query	0.128	0.145	-0.017
Webis	webis-t5-ax1	query	0.125	0.145	-0.019
Webis	webis-bm25-ax1	query	0.127	0.147	-0.020
Webis	webis-bm25-ax3	query	0.123	0.145	-0.021
Waterloo_Cormack	watgoogp	query	0.032	0.054	-0.021
UWaterlooMDS	baselineBM25	query	0.122	0.144	-0.022
Webis	webis-bm25	query	0.122	0.145	-0.023
Waterloo_Cormack	watrrfnp	query	0.048	0.072	-0.024
Waterloo_Cormack	watbm25	query	0.100	0.125	-0.025
Waterloo_Cormack	watrrfm	query	0.071	0.103	-0.032
Waterloo_Cormack	watbm25p	query	0.078	0.113	-0.035
Waterloo_Cormack	watrrfall	query	0.054	0.096	-0.042

Table 3: Automatic run results.

			Avg. Compatibility		
Group	Run	Topic Fields	help	harm	help-harm
h2oloo	vera_mt5_0.5	desc, stance	0.297	0.038	0.259
h2oloo	vera_mdt5_0.5	desc, stance	0.299	0.042	0.258
h2oloo	vera_mt5_0.95	desc, stance	0.298	0.053	0.245
h2oloo	vera_mdt5_0.95	desc, stance	0.294	0.065	0.228
h2oloo	vera0	desc, stance	0.247	0.031	0.216
UWaterlooMDS	WatSMC-Correct	query, desc, narr, stance, evidence	0.264	0.055	0.208
UWaterlooMDS	WatSMT-SD-S1	query, stance	0.224	0.037	0.187
h2oloo	mt5_r	desc, stance	0.267	0.093	0.174
UWaterlooMDS	WatSMC-CALQAHC2	query, desc, stance	0.222	0.055	0.168
UWaterlooMDS	WatSMM-CALQAAll	query, desc, stance	0.200	0.034	0.167
UWaterlooMDS	WatSMC-CALQAHC1	query, desc, stance	0.244	0.078	0.165
UWaterlooMDS	WatSMC-CALQAAll	query, desc, stance	0.225	0.080	0.146
UWaterlooMDS	WatSMM-CALQA100	query, desc, stance	0.206	0.064	0.142
UWaterlooMDS	WatSMT-SD-S2	query, stance	0.190	0.059	0.131
h2oloo	mdt5_r	desc, stance	0.251	0.128	0.122
CiTIUS	citius.r1	query, desc, stance	0.219	0.123	0.096
UWaterlooMDS	WatSMC-CALQA100	query, desc, stance	0.206	0.134	0.072
UWaterlooMDS	WatSMM-CAL	query	0.191	0.129	0.062
UWaterlooMDS	WatSMM-CALHC	query	0.198	0.136	0.062
UWaterlooMDS	WatSMM-Fused	query	0.201	0.140	0.061
CiTIUS	citius.r9	query, desc, stance	0.203	0.143	0.060
UWaterlooMDS	WatSMM-CALPR	query	0.195	0.139	0.056
UWaterlooMDS	WatSMC-CAL	query	0.222	0.167	0.055
CiTIUS	citius.r10	query, desc, stance	0.196	0.153	0.042
CiTIUS	citius.r5	query, desc, stance	0.194	0.159	0.035
CiTIUS	citius.r3	query, desc, stance	0.196	0.161	0.035
CiTIUS	citius.r2	query, desc, stance	0.188	0.166	0.022
CiTIUS	citius.r4	query, desc, stance	0.185	0.165	0.021
CiTIUS	citius.r6	query, desc, stance	0.185	0.166	0.019
CiTIUS	citius.r7	query, desc, stance	0.134	0.128	0.006

Table 4: Manual run results.

expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

References

[1] M. Abualsaud, M. D. Smucker, C. Lioma, M. Maistro, and G. Zuccon. Overview of the TREC 2019 Decision Track. In E. M. Voorhees and A. Ellis, editors, *The Twenty-Eigth Text REtrieval Conference Proceedings (TREC 2019)*. National Institute of Standards and Technology (NIST), Special Publication 1250, Washington, USA, 2020.

[2] M. Abualsaud, I. X. Chen, K. Ghajar, L. N. L. Minh, M. D. Smucker, A. V. Tahami, and D. Zhang. UWaterlooMDS at the TREC 2021 Health Misinformation Track. In I. Soboroff and A. Ellis, editors, *The Thirtieth REtrieval Conference Proceedings (TREC 2021)*. National Institute of Standards and Technology (NIST), Special Publication 500-335, Washington, USA, 2022.

[3] A. Bondarenko, M. Fröbe, M. Gohsen, S. Günther, J. Kiesel, J. Schwerter, S. Syed, M. Völske, M. Potthast, B. Stein, and M. Hagen. Webis at TREC 2021: Deep Learning, Health Misinformation, and Podcasts Tracks. In I. Soboroff and A. Ellis, editors, *The Thirtieth REtrieval Conference Proceedings (TREC 2021)*. National Institute of Standards and Technology (NIST), Special Publication 500-335, Washington, USA, 2022.

[4] C. L. A. Clarke, M. D. Smucker, and A. Vtyurina. Offline Evaluation by Maximum Similarity to an Ideal Ranking. In M. d'Aquin, S. Dietze, C. Hauff, E. Curry, and P. Cudré-Mauroux, editors, *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM 2020)*, pages 225–234. ACM, New York, USA, 2020.

[5] C. L. A. Clarke, A. Vtyurina, and M. D. Smucker. Offline Evaluation without Gain. In K. Balog, V. Setty, C. Lioma, Y. Liu, M. Zhang, and K. Berberich, editors, *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR 2020)*, pages 185–192. ACM, New York, USA, 2020.

[6] C. L. A. Clarke, , M. M. S. Rizvi, M. D. Smucker, and G. Zuccon. Overview of the TREC 2020 Misinformation Track. In E. M. Voorhees and A. Ellis, editors, *The Twenty-Ninth Text REtrieval Conference Proceedings (TREC 2020)*. National Institute of Standards and Technology (NIST), Special Publication 1266, Washington, USA, 2021.

[7] M. Fernandez-Pichel, M. Prada-Corral, D. E. Losada, J. C. Pichel, and P. Gamallo. CiTIUS at the TREC 2021 Health Misinformation Track. In I. Soboroff and A. Ellis, editors, *The Thirtieth REtrieval Conference Proceedings (TREC 2021)*. National Institute of Standards and Technology (NIST), Special Publication 500-335, Washington, USA, 2022.

[8] C. Lioma, J. G. Simonsen, and B. Larsen. Evaluation Measures for Relevance and Credibility in Ranked Lists. In J. Kamps, E. Kanoulas, M. de Rijke, H. Fang, and E. Yilmaz, editors, *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR 2017)*, pages 91–98. ACM, New York, USA, 2017.

[9] J. Palotti, G. Zuccon, and A. Hanbury. MM: A new Framework for Multidimensional Evaluation of Search Engines. In A. Cuzzocrea, J. Allan, N. W. Paton, D. Srivastava, R. Agrawal,

A. Z. Broder, M. J. Zaki, K. S. Candan, A. Labrinidis, A. Schuster, and H. Wang, editors, *Proceedings of the 27th ACM International Conference on Information & Knowledge Management (CIKM 2018)*, pages 1699–1702. ACM, New York, USA, 2018.

[10] R. Pradeep, X. Ma, R. Nogueira, and J. Lin. Vera: Prediction Techniques for Reducing Harmful Misinformation in Consumer Health Search. In F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, and T. Sakai, editors, *The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR 2021)*, pages 2066–2070. ACM, New York, USA, 2021.

[11] I. B. Schlicht, A. F. Magnossao de Paula, and P. Rosso. UPV at TREC Health Misinformation Track 2021 Ranking with SBERT and Quality Estimators. In I. Soboroff and A. Ellis, editors, *The Thirtieth REtrieval Conference Proceedings (TREC 2021)*. National Institute of Standards and Technology (NIST), Special Publication 500-335, Washington, USA, 2022.

[12] B. Zhang, N. Naderi, F. Jaume-Santero, and D. Teodoro. DS4DH at TREC Health Misinformation 2021: Multi-Dimensional Ranking Models with Transfer Learning and Rank Fusion. In I. Soboroff and A. Ellis, editors, *The Thirtieth REtrieval Conference Proceedings (TREC 2021)*. National Institute of Standards and Technology (NIST), Special Publication 500-335, Washington, USA, 2022.