Overview of the TREC 2022 Health Misinformation Track (Notebook)

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

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

Abstract

This notebook version of the overview is sparse on details. It's primary purpose is to make available to all participants the overall results. The final overview will be updated with more details specific to the 2022 track.

1 Introduction

TREC 2022 was the fourth and final year for the Health Misinformation track, which was named the Decision Track in 2019 [1]. In 2022, the track had an answer prediction task as well as a web retrieval task. In each year, the track has used a crawl for its document collection. In 2019, 2021, and 2022 we used web crawls, and in 2020, we used a web crawl restricted to news sites.

By focusing on health-related 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 and quality.

This year, each topic was expressed as a yes/no 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. Based on a credible source of information, we declare an *answer* for a topic as either *yes* or *no*. We provide an *evidence* URL link to the source we used to determine the stance. Each topic is also supplied with a background providing additional clarification to the assessors. We did not provide the answers and evidence to participants until after the evaluation. Automatic runs could only make use of the topic's question or query.

Answer prediction runs consisted of both a prediction of yes or no for a topic's question, and also a numeric score ranging from 0 to 1 where 1 was "yes". Using the numeric score, prediction runs were evaluated based on their AUC. We also report the runs' true positive rate (TPR), false positive rate (FPR), and accuracy given their predicted label/answer. The positive class is "yes".

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 retrieval runs are evaluated based on their *compatibility* [2, 3] 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/quality of the documents.

2 Topics

We created 50 topics this year with half of them having an answer of yes and half with an answer of no. NIST was only able to provide assessments for 45 of the 50 topics. Of these 45 topics, no harmful documents were found for topics 165, 171, 174, 176, 180, 182, 191, and 200. We have excluded these eight topics from the analysis in this paper.

```
<topic>
<number>155</number>
<question>Can you use WD-40 for arthritis?</question>
<query>WD-40 arthritis</query>
<background>WD-40 is an oil-based lubricant. Arthritis is
a health condition where the joints are inflamed and swollen,
with symptoms such as joint pain and stiffness.  This question
is asking if an arthritis sufferer could get relief from the
pain by rubbing WD-40 on their joints.</background>
<disclaimer>We do not claim to be providing medical advice,
and medical decisions should never be made based on the answer
we have chosen. Consult a medical doctor for professional
advice.</disclaimer>
<answer>no</answer>
<evidence>https://wd40.co.uk/tips-and-tricks/
can-wd-40-help-arthritis-stiff-joints/</evidence>
</topic>
```

Figure 1: Example of a topic for the TREC 2022 Health Misinformation track.

3 Document Collection

This year we again 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 Evaluation

Retrieval runs were evaluated by using a script[2] to compute the compatibility measure [2, 3]. We derive a qrels file to use with compatibility from the original NIST qrels file and preference judgments of the assessors collected for top documents.

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

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

4.0.1 Preference Levels

For the compatibility measure, we converted the 2 aspects judged for documents (usefulness, answer) into a basic preference ordering where very-useful documents are preferred to useful documents, and correct documents are preferred to unclear documents. Correct and unclear documents are preferred to not-useful documents. Not-useful documents are preferred to incorrect documents. Of the incorrect documents, the very-useful, incorrect documents are least preferred. Documents judged to be *very-useful* with a correct answer were then preference judged to find up to the top 10 preferred documents. For some topics, *useful* documents were also preference judged. As we proceeded with preference judging, we found it important to limit the pool of documents for preference judging, and we restricted the documents to correct very-useful documents.

We use the preference ordering to create a set of helpful and harmful preference qrels. 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.

4.1 Evaluation Measures

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

5 Results

Tables 1 and 2 report the results for prediction and retrieval runs. Figure 1 shows ROC curves for each group's top automatic and manual prediction runs. Figure 2 shows the harmful compatibility of retrieval runs plotted against helpful compatibility. 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.

6 Acknowledgments

Thanks to Linh Nhi Phan Minh for help with topic creation, testing, and processing of data for use with the judging system. Thanks to Amir Vakili Tahami and Dake Zhang for creating a nicely formatted version of the document collection, testing, and analysis of results. Thanks to Kamyar Ghajar for help with processing MS-Marco for topic creation.

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 the Digital Research Alliance of 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 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] 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,

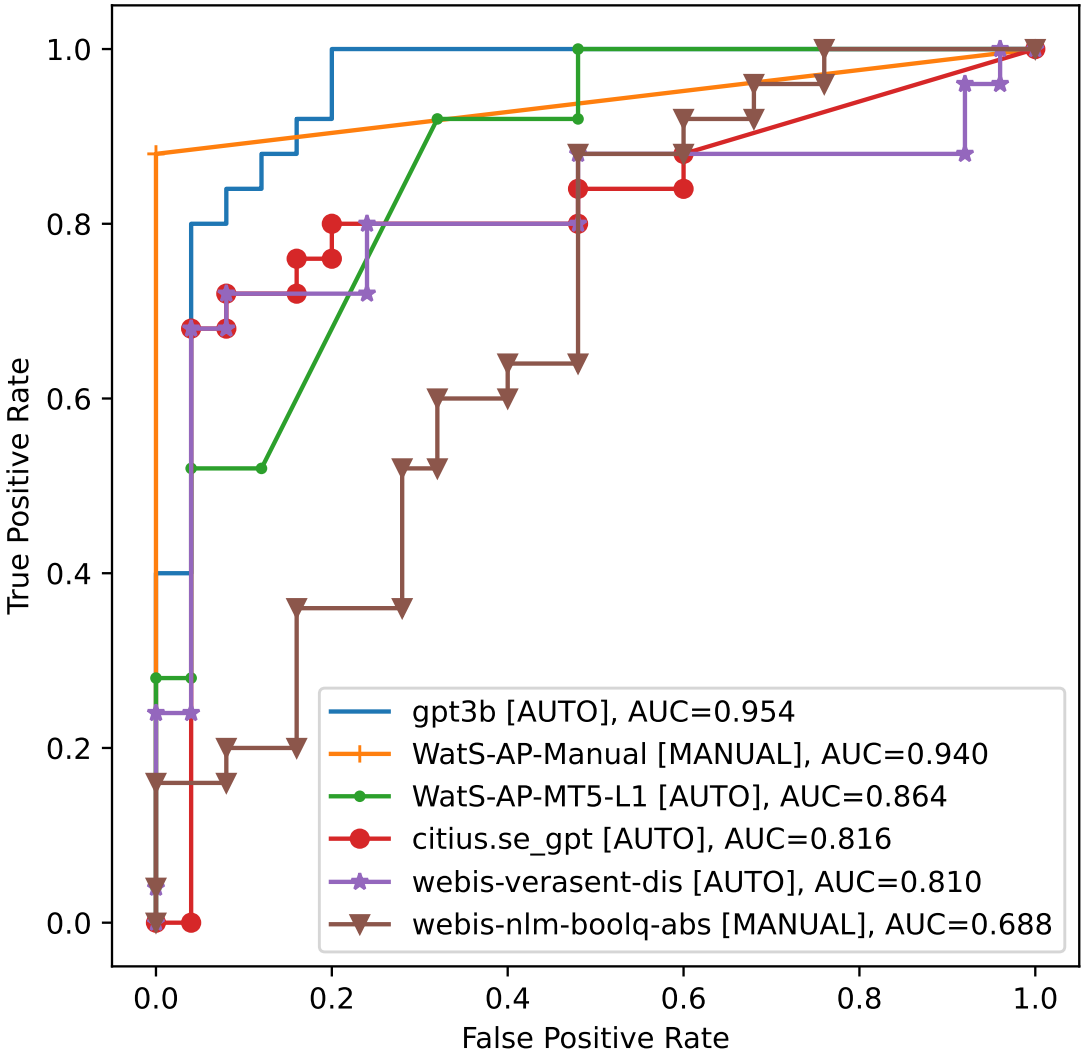

Figure 2: Receiver operating characteristic (ROC) curves and area under the curve (AUC) of groups' top automatic and manual answer prediction runs.

Group	Run	Type	Accuracy	TPR	FPR	AUC
h2oloo	gpt3b	auto	0.86	0.76	0.04	0.954
h2oloo	gpt3a	auto	0.86	0.76	0.04	0.952
UWaterlooMDS	WatS-AP-Manual	manual	0.94	0.88	0.00	0.940
h2oloo	vera_gpt3	auto	0.88	0.80	0.04	0.934
h2oloo	vera_gpt3_abs	auto	0.88	0.80	0.04	0.880
UWaterlooMDS	WatS-AP-MT5-L1	auto	0.70	1.00	0.60	0.864
h2oloo	gpt3a_fc	auto	0.86	0.76	0.04	0.860
UWaterlooMDS	WatS-manual-pred	manual	0.84	0.76	0.08	0.840
h2oloo	vera	auto	0.68	0.84	0.48	0.821
CiTIUS	citius.se_gpt	auto	0.80	0.68	0.08	0.816
UWaterlooMDS	WatS-AP-MT5	auto	0.64	0.96	0.68	0.813
Webis	webis-verasent-dis	auto	0.70	0.80	0.40	0.810
Webis	webis-longck-dis	auto	0.64	0.64	0.36	0.790
CiTIUS	citius.gpt-3	auto	0.76	0.68	0.16	0.767
CiTIUS	citius.se	auto	0.62	0.96	0.72	0.707
UWaterlooMDS	WatS-BB75-MT5-TA	auto	0.66	0.44	0.12	0.691
Webis	webis-nlm-boolq-abs	manual	0.52	1.00	0.96	0.688
h2oloo	vera_abs	auto	0.68	0.84	0.48	0.680
Webis	webis-longck-uniqa-dis	auto	0.62	0.72	0.48	0.664
Webis	webis-uniqa-dis	auto	0.62	0.72	0.48	0.659
Webis	webis-longck-uniqa-ax-dis	auto	0.60	0.68	0.48	0.656
Webis	webis-goo-boolq-abs	auto	0.52	1.00	0.96	0.653
UWaterlooMDS	WatS-AP-Baseline-L1	auto	0.54	0.72	0.64	0.565
UWaterlooMDS	WatS-AP-Baseline	auto	0.56	0.80	0.68	0.557
Webis	webis-goo-lbert-title-abs	auto	0.50	0.92	0.92	0.483
Webis	webis-nlm-lbert-abs	manual	0.50	0.80	0.80	0.483
Webis	webis-goo-lbert-abs	auto	0.50	0.88	0.88	0.478
h2oloo	gpt3a_neg	auto	0.14	0.24	0.96	0.048
h2oloo	gpt3b_neg	auto	0.14	0.24	0.96	0.046

Table 1: Answer prediction runs. Reported are a run's accuracy, true positive rate (TPR), false positive rate (FPR), and area under the ROC curve (AUC). AUC is the primary measure. All questions were yes/no questions, and "yes" is the positive class.

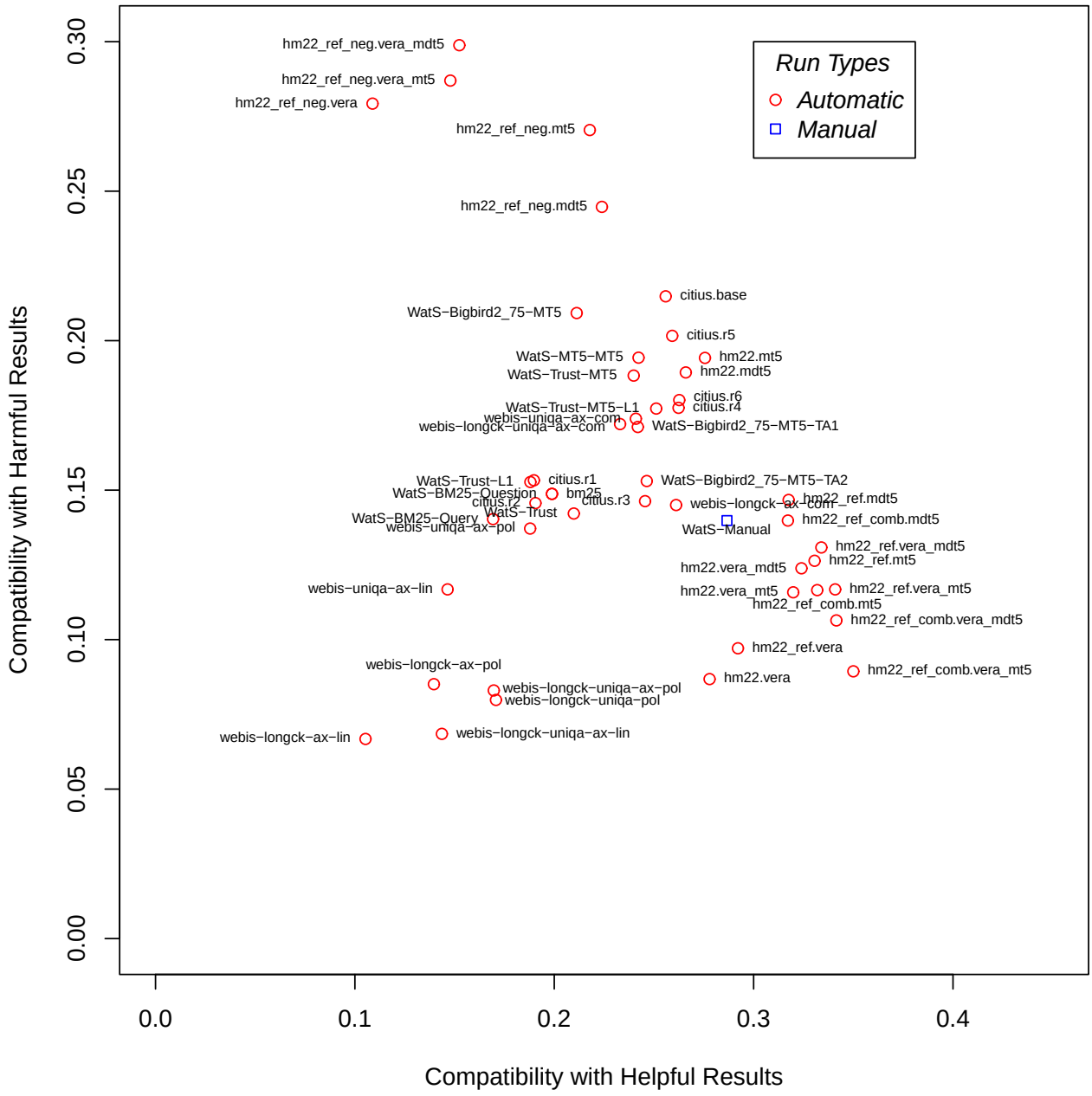

Figure 3: Compatibility of runs with helpful and harmful results. A good run is helpful and not harmful. For a given level of helpfulness, a run with less harm is to be preferred.

			Avg. Compatibility		
Group	Run	Type	help	harm	help-harm
h2oloo	hm22_ref_comb.vera_mt5	auto	0.350	0.089	0.261
h2oloo	hm22_ref_comb.vera_mdt5	auto	0.342	0.106	0.235
h2oloo	hm22_ref.vera_mt5	auto	0.341	0.117	0.224
h2oloo	hm22_ref_comb.mt5	auto	0.332	0.117	0.215
h2oloo	hm22_ref.mt5	auto	0.331	0.126	0.204
h2oloo	hm22.vera_mt5	auto	0.320	0.116	0.204
h2oloo	hm22_ref.vera_mdt5	auto	0.334	0.131	0.203
h2oloo	hm22.vera_mdt5	auto	0.324	0.124	0.200
h2oloo	hm22_ref.vera	auto	0.292	0.097	0.195
h2oloo	hm22.vera	auto	0.278	0.087	0.191
h2oloo	hm22_ref_comb.mdt5	auto	0.317	0.140	0.177
h2oloo	hm22_ref.mdt5	auto	0.318	0.147	0.171
UWaterlooMDS	WatS-Manual	manual	0.287	0.140	0.147
Webis	webis-longck-ax-com	auto	0.261	0.145	0.116
CiTIUS	citius.r3	auto	0.246	0.146	0.099
UWaterlooMDS	WatS-Bigbird2_75-MT5-TA2	auto	0.246	0.153	0.093
Webis	webis-longck-uniqa-pol	auto	0.171	0.080	0.091
Webis	webis-longck-uniqa-ax-pol	auto	0.170	0.083	0.087
CiTIUS	citius.r4	auto	0.262	0.178	0.085
CiTIUS	citius.r6	auto	0.263	0.180	0.083
h2oloo	hm22.mt5	auto	0.276	0.194	0.081
h2oloo	hm22.mdt5	auto	0.266	0.189	0.077
Webis	webis-longck-uniqa-ax-lin	auto	0.144	0.069	0.075
UWaterlooMDS	WatS-Trust-MT5-L1	auto	0.251	0.177	0.074
UWaterlooMDS	WatS-Bigbird2_75-MT5-TA1	auto	0.242	0.171	0.071
UWaterlooMDS	WatS-Trust	auto	0.210	0.142	0.068
Webis	webis-uniqa-ax-com	auto	0.241	0.174	0.067
Webis	webis-longck-uniqa-ax-com	auto	0.233	0.172	0.061
CiTIUS	citius.r5	auto	0.259	0.202	0.058
Webis	webis-longck-ax-pol	auto	0.140	0.085	0.054
UWaterlooMDS	WatS-Trust-MT5	auto	0.240	0.188	0.052
Webis	webis-uniqa-ax-pol	auto	0.188	0.137	0.051
h2oloo	bm25	auto	0.199	0.149	0.050
UWaterlooMDS	WatS-BM25-Question	auto	0.199	0.149	0.050
UWaterlooMDS	WatS-MT5-MT5	auto	0.242	0.194	0.048
CiTIUS	citius.r2	auto	0.191	0.146	0.045
CiTIUS	citius.base	auto	0.256	0.215	0.041
Webis	webis-longck-ax-lin	auto	0.105	0.067	0.038
CiTIUS	citius.r1	auto	0.190	0.153	0.037
UWaterlooMDS	WatS-Trust-L1	auto	0.188	0.153	0.035
Webis	webis-uniqa-ax-lin	auto	0.146	0.117	0.030
UWaterlooMDS	WatS-BM25-Query	auto	0.169	0.140	0.029
UWaterlooMDS	WatS-Bigbird2_75-MT5	auto	0.211	0.209	0.002
h2oloo	hm22_ref_neg.mdt5	auto	0.224	0.245	-0.021
h2oloo	hm22_ref_neg.mt5	auto	0.218	0.270	-0.053
h2oloo	hm22_ref_neg.vera_mt5	auto	0.148	0.287	-0.139
h2oloo	hm22_ref_neg.vera_mdt5	auto	0.152	0.299	-0.146
h2oloo	hm22_ref_neg.vera	auto	0.109	0.279	-0.170

Table 2: Retrieval runs. The primary measure is compatibility-help minus compatibility-harm (help-harm).

editors, *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM 2020)*, pages 225–234. ACM, New York, USA, 2020.

[3] 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.