VOH.CoLAB at TREC 2020 Health Misinformation Track*

Simão N. Gonçalves[1,2] and Flávio Martins[1,2,3]

[1] VOH.CoLAB, Universidade NOVA de Lisboa, Lisboa, Portugal
[2] NOVA LINCS, Universidade NOVA de Lisboa, Caparica, Portugal
[3] Comprehensive Health Research Centre (CHRC), Lisboa, Portugal
simao.goncalves@vohcolab.org flavio.martins@vohcolab.org

Abstract. In this paper, we describe the participation of VOH.CoLAB in the TREC 2020 Health Misinformation Track (HMT). This year's edition of the track focused on two main *Consumer Health Search* tasks regarding COVID-19 questions: 1) to find misinformation; 2) to find relevant, credible, and correct information. In our participation in the HMT track, we submitted runs to both tasks, performing experiments to explore two main research hypothesis: 1) Does misinformation avoid mentioning the evidence text? 2) Does correct and credible information look similar to the evidence text? To explore these two complementary ideas we represent both the documents and the evidence as vectors and compute scores using a formula based on Kullback-Leibler divergence.

Keywords: Consumer Health Search · COVID-19 · Misinformation.

1 Introduction

The outbreak of a novel coronavirus in 2019 (SARS-CoV-2), has led to an increase of misinformation online and pushed the World Health Organization (WHO) to declare "We're fighting an infodemic". In *Consumer Health Search*, people often cannot separate the useful, correct, and credible information from misinformation. Therefore, because *Consumer Health Search* is used to make decisions about Health, fighting misinformation is critical. Previous research [2] showed that users make incorrect decisions when presented increasing amounts of incorrect information in Search Engine Result Pages (SERPs), which is potentially harmful.

The TREC Health Misinformation Track focuses on methods that promote correct and credible information over non-credible, incorrect information. Access to correct information about COVID-19 is of critical importance to public health, therefore this track aims to create a suitable dataset to foster research on misinformation, credibility, correctness, and relevance regarding COVID-19 questions. A dataset labeled along these dimensions can be used to build information retrieval systems that retrieve correct information and demote misinformation. This can lead people to make better decisions about their health.

* This work is supported by DSAIPA project FrailCare.AI (DSAIPA/0106/2019/02) and by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT – Fundação para a Ciência e a Tecnologia, through national funds.

2 Health Misinformation Track Setup

This section describes the setup of the track, which aims to obtain a labeled dataset with annotations for document usefulness, correctness, and credibility suitable for future research in health misinformation.

2.1 Collection

The collection contains 65 million news articles from CommonCrawl News[4] corresponding to the period from January to April of 2020. CommonCrawl News contains archives of Web pages (snapshots) crawled from news sites from all over the world in multiple languages and published by the CommonCrawl Foundation.

2.2 Topics

The track focuses on *Consumer Health Search*, where the users are regular people seeking health advice online in regards to the COVID-19 epidemic. This represents the dangers of the proliferation of misinformation and the weaknesses of search engines, which can have negative consequences on consumer health. A total of 50 topics were developed for evaluation on this track with the following fields:

number: id of the topic | e.g., 16

title: a pair of *treatment* and *disease* | e.g., Vinegar COVID-19

description: a query in the form of "Can *treatment effect* COVID-19?". The effect is either one of the following: cause, prevent, worsen, cure or help. | e.g., Can vinegar prevent COVID-19?

answer: answer provided by the writer's best understanding at the time of writing the topic. | e.g., no

evidence: URL of a page from the open Web that was used to determine this answer. | e.g., `https://globalnews.ca/news/6703882/coronavirus-covid-19-cleaning-vinegar/`

narrative: Extract from evidence URL which supports the answer. | e.g.'Vinegar is known to have disinfectant properties. Recently, vinegar has been promoted as a disinfectant (...)'

[4] `https://commoncrawl.org/2016/10/news-dataset-available/`

2.3 Tasks

The Health Misinformation Track consist of two tasks:

Total Recall Task: find misinformation articles contradicting the answer.

***Ad hoc* Task**: find useful, correct, and credible articles supporting the answer.

While in the Total Recall task the goal is to find harmful misinformation, the *Ad Hoc* task aims to find relevant, credible, and correct information that can help consumer health searchers make better health-related decisions.

2.4 Evaluation

Relevance judgments were evaluated according to three criteria: *usefulness*, *correctness*, and *credibility*. A *useful* document talks about the topic's subject, but doesn't depend on correctness (effectively it's the "topic relevance"). A *correct* document must give the right answer to the topic (therefore it is also useful). Finally, a *credible* document doesn't depend on correctness, but must be useful. Credibility judgments also depend on other factors such as: amount of expertise of the author, references contained to support claims, whether the article contains advertising, and more. Further information on how the annotators classified the relevance judgments can be found in the track's Assessing Guidelines [5].

Derived qrels are special qrel files, derived from the 2020 qrels, that were used to evaluate adhoc and total recall runs on multiple levels. The derived qrels are essentially combinations of all the three criteria considered in the relevance judgments. In the Results section we discuss more in depth these derived qrels.

Total Recall is evaluated according to the metric *Rprec* of the incorrect documents retrieved. Incorrect documents are useful, but don't give the right answer to the topic.

AdHoc Retrieval is evaluated for 9 different criteria such as: finding only useful documents, or finding useful, correct, and credible documents, among other combinations of these three aspects. The metrics used for this task were: Normalized Discounted Cumulative Gain (NDCG), Convex Aggregating Measure (CAM) [4] using Mean Average Precision (MAP), and *compatibility* [1].

[5] `https://trec-health-misinfo.github.io/docs/AssessingGuidelines-2020.pdf`

3 Experimental methodology

In this section we describe the experimental methodology including indexing and data preprocessing.

3.1 Preprocessing

The first step filters out non-English documents since these are considered non-relevant according to the assessors' instructions. To create a whitelist of English pages we adapted the code from *CC-News Tools*[6]. The final indexing contains 28.3M documents mostly in English, compared to the original 65M documents.

3.2 Indexing and Retrieval

We used **Anserini** [6] to create the index, using its *whitelist* flag to index only English documents. **Pyserini**[7] was used to query the index of the collection using queries composed by the concatenation of the topic's fields *title* and *description*, because some of the topics' *title* field contained extra information not present in the *description* field. For our re-ranking runs we use **nltk**[8] to preprocess and tokenize the documents' text and the Python package **scikit-learn** [5] to create a tf-idf representation of the candidate documents retrieved initially.

4 Submitted runs

This section describes our approach at answering both tasks described above. In summary, we have submitted a total of 5 runs, they are:

Total Recall
- vohbm25rm3 (baseline)
- vohEvDivTfidf
- vohEvDiv_colm

***Ad hoc* Retrieval**
- vohbm25 (baseline)
- vohcolabEvSim

4.1 Baselines

Total Recall Task. Our baseline uses BM25 and RM3 to retrieve 10k documents. This task is recall-oriented and therefore we employ pseudo-relevance feedback:

1. First, we use BM25 to retrieve the most likely relevant documents;
2. Then, we assume blindly that the top documents retrieved are relevant to the topic and use RM3 to expand the query to find more relevant documents.

Ad hoc Task. Our baseline uses BM25 alone to retrieve 1000 documents. We do not employ query expansion as this task is precision-oriented.

[6] `https://github.com/jmmackenzie/cc-news-tools/`

[7] `https://github.com/castorini/pyserini`

[8] `https://www.nltk.org`

4.2 Commons similarity and divergence

This section contains three submissions that arise from the following hypothesis: "Misinformation tries to avoid the actual evidence, therefore vocabulary distribution will be different". We further make an analogous argument for the *Ad hoc* Retrieval task: "Correct information paraphrases the actual evidence, therefore vocabulary distribution will be similar". To test this hypothesis we built the following pipeline for each topic:

1. Retrieve the evidence text of the topic by crawling the evidence URL field;
2. Retrieve an initial set of documents with BM25;
3. Represent documents as probability distributions;
4. Rerank on the similarity between documents and the evidence text.

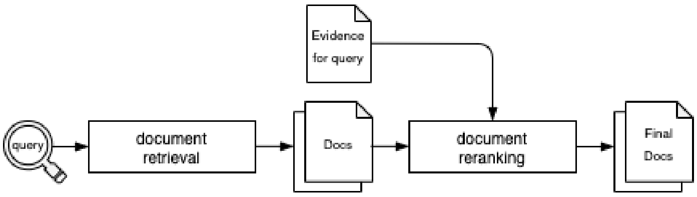


Fig. 1. KL reranking pipeline

Fig 1 shows the core pipeline of the three remaining submissions. The settings for each run can be found in Table 1.

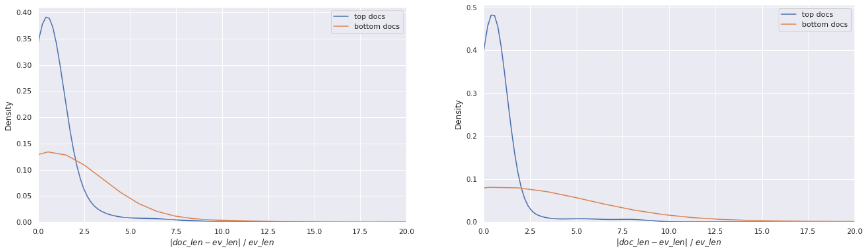
In order to represent documents as probability distributions in vector space we create tf-idf vectors of the initial retrieval and normalize the document vectors, so that they add up to 1. To compute the distance between each document and the evidence we propose Equation (1). This is a symmetrized version of the Kullback-Leibler divergence and was adapted from the work of Kulkarni and Callan [3] where it was originally used to cluster topics in a collection based on the distance between each document and their centroid.

$$\begin{aligned} D_{NKL}\left(\theta_E \mid\mid \theta_D\right) = & \sum_{w \in D \cap Q} p(w \mid \theta_E) \log \frac{p(w \mid \theta_D)}{\lambda \cdot p(w \mid \theta_B)} \\ & + \sum_{w \in D \cap Q} p(w \mid \theta_D) \log \frac{p(w \mid \theta_E)}{\lambda \cdot p(w \mid \theta_B)} \end{aligned} \tag{1}$$

We adapt their formula to our scenario by computing the distance between the evidence text document θ_E and each document present in the initial retrieval θ_{D_i}. We also replace the global collection model θ_C by a topic-based background model θ_B, which is estimated using an average pooling of the documents retrieved by a query Q for each given topic T.

Table 1. Pipeline description for commons similarity and divergence runs.

Run	matrix	normalization	smoothing	λ	Distance
vohEvDivTfidf	TF-IDF	document-level	No	0.3	Neg-Sym KL
vohEvDiv_colm	TF	term-level	No	0.3	Neg-Sym KL
vohcolabEvSim	TF-IDF	document-level	Yes	N/A	KL



(a) vohbm25 (b) vohEvDivTfidf

Fig. 2. Distribution of the difference of doc-length's between top/bottom vs. evidence.

Document length bias We suspected that documents with a length similar to the evidence text's have an advantage due to the use of the *KL-divergence*. Documents larger than the evidence text probably contain a larger vocabulary, thus increasing the divergence, while smaller documents probably contain a smaller vocabulary than the evidence text, thus also contributing to an increase in the divergence. We confirmed this bias by comparing the distribution of document lengths in BM25 retrieval in Fig. 2a with *vohEvDivTfidf* in Fig. 2b. As expected, in *vohEvDivTfidf* the top documents are concentrated closer to the evidence text's size.

5 Results

In this section we look at the results obtained in both the submitted runs and not-submitted runs. We discuss the experiments evaluated on each task.

Table 2. Total Recall Task results

Run	Rprec
median	0.0976
vohbm25rm3	**0.1026**
vohTR_bm25*	0.1020
vohbm25rm3-ml*	0.0482
vohEvDiv_colm	0.0430
vohEvDivTfidf	0.0325

5.1 Total Recall Results

The goal in the *Total Recall* task is to find incorrect documents. Table 2 shows the results obtained for this task. We include the Track leaderboard's median score, our submitted runs, and also non-submitted runs (identifiable by an asterisk).

The baseline *vohbm25rm3* was slightly better than the median at finding incorrect documents. We also ran the baseline on the index containing the original, unfiltered collection of multi-language articles.

5.2 AdHoc Results

The *AdHoc* task has relevance judgments for 3 criteria: usefulness, credibility, and correctness. Based on these criteria, several derived relevance judgments were used for evaluation. We grouped the measured relevance judgments in two: the binary assessed judgments in Table 3, and the multi-aspect assessed judgments in Table 4. The multi-aspect measures are computed using CAM and *compatibility*.

Similarly to the Total Recall task we also ran the baseline over the original, unfiltered, multi-language collection. The submitted run *vohcolabEvSim* had the document rank inverted by accident, so we also evaluated the not-submitted run *vohcolabEvSimInv** with the scores corrected.

Table 3. *Ad hoc* Task results using binary relevance judgments.

	NDCG			
	useful	useful correct	useful credible	useful correct credible
median	0.4699	0.3380	0.4471	0.3308
vohbm25	**0.6077**	**0.4771**	**0.5768**	**0.4592**
vohAH_bm25rm3*	0.5950	0.4564	0.5684	0.4545
vohcolabEvSimInv*	0.5528	0.4547	0.5504	0.4483
vohbm25-ml*	0.4800	0.3408	0.4608	0.3317
vohcolabEvSim	0.4287	0.3271	0.3950	0.3061

Table 4. *Ad hoc* Task results using multiple aspect judgments.

	cam_map			compatibility		
	useful credible	correct credible	3aspects	helpful-only	harmful-only	helpful-harmful
median	0.1717	0.1003	0.1389	0.3337	0.0747	0.2590
vohbm25	**0.2824**	**0.1740**	**0.2367**	**0.3402**	**0.1124**	0.2278
vohAH_bm25rm3*	0.2929	0.1891	0.2468	0.3589	0.1218	**0.2371**
vohcolabEvSimInv*	0.2023	0.1486	0.1742	0.2454	0.0457	0.1997
vohbm25-ml*	0.1869	0.0952	0.1514	0.2293	0.0503	0.1790
vohcolabEvSim	0.0717	0.0330	0.0574	0.0862	0.0245	0.0608

5.3 Conclusions

The baselines outperformed the median by a significant margin and were the best performing runs for both tasks. We also showed that using an English filtered index helped find more relevant documents for both tasks. RM3 improved the compatibility *compatibility* measures by a significant margin on the *adhoc* task.

While the officially submitted run *vohcolabEvSim* based on the *commons similarity/divergence hypothesis* under-performed due to the scores being inverted, the post-mortem run *vohcolabEvSimInv**, which inverts the scores back, achieved much better results in the AdHoc task. It significantly increased the *compatibility* with helpful-only results and keeps a low *compatibility* with harmful-only results.

References

1. Clarke, C.L., Vtyurina, A., Smucker, M.D.: Offline evaluation without gain. In: Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval. p. 185–192. ICTIR '20, Association for Computing Machinery, New York, NY, USA (2020), `https://doi.org/10.1145/3409256.3409816`
2. Islam, M.S., Sarkar, T., Khan, S.H., Mostofa Kamal, A.H., Hasan, S.M.M., Kabir, A., Yeasmin, D., Islam, M.A., Amin Chowdhury, K.I., Anwar, K.S., Chughtai, A.A., Seale, H.: Covid-19–related infodemic and its impact on public health: A global social media analysis. The American Journal of Tropical Medicine and Hygiene (2020), `http://www.ajtmh.org/content/journals/10.4269/ajtmh.20-0812`
3. Kulkarni, A., Callan, J.: Selective search: Efficient and effective search of large textual collections. ACM Trans. Inf. Syst. 33(4) (Apr 2015), `https://doi.org/10.1145/2738035`
4. Lioma, C., Simonsen, J.G., Larsen, B.: Evaluation measures for relevance and credibility in ranked lists. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval. p. 91–98. ICTIR '17, Association for Computing Machinery, New York, NY, USA (2017), 2
5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
6. Yang, P., Fang, H., Lin, J.: Anserini: Enabling the Use of Lucene for Information Retrieval Research. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1253–1256. SIGIR '17, ACM, New York, NY, USA (2017)