UPV at TREC Health Misinformation Track 2021
Ranking with SBERT and Quality Estimators

Ipek Baris Schlicht and **Angel Felipe Magnossão de Paula** and **Paolo Rosso**
Universitat Politècnica de València, Spain

Abstract

Health misinformation on search engines is a significant problem that could negatively affect individuals or public health. To mitigate the problem, TREC organizes a health misinformation track. This paper presents our submissions to this track. We use a BM25 and a domain-specific semantic search engine for retrieving initial documents. Later, we examine a health news schema for quality assessment and apply it to re-rank documents. Finally, we merge the scores from the different components by using reciprocal rank fusion. Finally, we discuss the results and conclude with future works.

1 Introduction

People widely use Web search engines to seek health information and get medical advice on their health conditions (Gualtieri, 2009). However, the Web hosts health misinformation. The presence of health misinformation could have adverse effects on individuals who believe in everything they read. To address this issue, this year, the TREC 2021 organized a shared task [1]. The task aims to implement an information retrieval that promotes helpful and credible information. An ideal search engine ranks credible and useful documents at the top of the ranking and ignores harmful documents.

Health articles could present multiple aspects of medical research. Thus, the assessment of health articles is beyond fact-checking. Some medical experts and journalists develop schemes to assess the quality indicators of a health article. DISCERN (Charnock et al., 1999) and the criteria from the Health News Review[2] are the well-known ones. Also, health articles could contain medical terms; thus, the language models trained on the general domain would not be sufficient to encode knowledge in the texts. In this study, we exploit the criteria from the Health News Review, select the top 4 criteria that the transformers (Vaswani et al., 2017) could perform well, and fine-tune a RoBERTa (Liu et al., 2019) classifier on the Health News Review dataset. We determine a reference vector that satisfies these criteria and then measure the distance between the reference vector and quality vector computed from each document using cosine similarity. Then we use the scores for ranking the documents retrieved by a BM25 model. We also implement a semantic search engine based on domain-specific sentence transformer. Finally, we merge the ranks from the different components to get fused rank lists.

2 Task

The input of the task is a topic which is about a health issue and medical treatment. The participants develop search engines that ideally return credible and correct information at the top of the ranking and discard incorrect information about a topic. The shared task dataset comprises two corpora: one for indexing, one for the topics. The dataset for indexing is a subset of the C4 corpus used by Google to train their T5 model[3].

The NIST assessors derived query-relevance files (qrels) based on usefulness, correctness, and credibility scores (Clarke et al., 2020a) by using 35 topics. Briefly, these criteria are:

- Usefulness measures how much a user finds an answer useful.
- Correctness is computed by checking whether the answer matches the useful document.

[1] https://trec-health-misinfo.github.io/, accessed on 27.10.2021

[2] https://www.healthnewsreview.org/about-us/review-criteria/, accessedon27.10.2021

[3] https://www.tensorflow.org/datasets/catalog/c4, accessed on 27.10.2021

TREC ID	BM25	SBERT	QE$^{\heartsuit}$	QE$^{\diamond}$
upv_bm25	■			
upv_fuse_2	■	■		
upv_fuse_3	■		■	
upv_fuse_5	■			■
upv_fuse_7	■	■	■	
upv_fuse_9	■	■		■

Table 1: QE denotes query estimation. QE$^{\heartsuit}$ uses RoBERTa-base and QE$^{\diamond}$ uses RoBERTa-large.

- Credibility measures how credible the document is.

3 Methodology

We describe our submissions[4] to the track as can be seen in Table 1. Initially, we used two search engines. The first one is BM25 (Robertson and Zaragoza, 2009) and the other one is based on a domain-specific Sentence-BERT (SBERT) (Reimers and Gurevych, 2019). Due to the limited resources, we indexed the search engines with the subset of the TREC corpus. This index was composed of 10k documents for each query, which were retrieved by the Pyserini implementation of BM25 (Yilmaz et al., 2020; Lin et al., 2021). Afterward, the ranks from the BM25 search engine were reranked with each quality estimator. In the end, in order to get the final ranks, we fused the ranks of BM25, SBERT, and the quality estimators. The following subsections describe the details of the submissions.

3.1 Search Engines

BM25 is a traditional retrieval method that has been widely used as a baseline retrieval system. We used the Pyserini implementation of the BM25 search engine to obtain the initial top 1000 documents for each topic.

In addition to the BM25, we also performed a semantic search. In the semantic search, the query and the documents are embedded with an SBERT model, and then based on the cosine similarity of the embeddings, a rank list is returned for each query. Since the health articles may contain medical terms and the standard transformer models such as BERT may not be representative for these documents, we used a model pre-trained on the PubMed articles (Lee et al., 2020). However, the model could encode only the limited number of tokens in a text. For this reason, prior to the encoding process, we first identified sentences in a document and then encoded only the first 20 sentences with the SBERT.

[4]We submitted officially ten runs to TREC, we noticed that 4 of them which use a Kullback-Leibler divergence score, have an implementation issue in reranking. Therefore, we only presented the other runs in this paper.

3.2 Quality Estimators for Reranking

A quality estimator is a multi-label classifier that gives a probabilistic score for each criterion. To implement a set of quality estimators, we used a publicly available dataset constructed from the Health News Review (Zuo et al., 2021). According to the dataset paper, RoBERTa-based models have shown promising results on criteria 1, 2, 7, and 8 (see Table 2). For this reason, we fine-tuned base and large RoBERTa models (Liu et al., 2019) by using the Huggingface library (Wolf et al., 2020) and used them to label unseen samples based on these criteria.

We hypothesize that an ideal health article would fulfill the quality criterion. Thus, we assumed a reference vector for an ideal article has 1.0 as a probability score for each criterion. Then, for each document in a topic, we estimated the quality criterion using the quality estimators and then measured the similarity score between the reference vector by calculating the cosine similarity. Finally, the documents were ranked again based on the similarity score. We repeated this process for the RoBERTa-base (QE$^{\heartsuit}$) and RoBERTa-large (QE$^{\diamond}$) models, and we obtained two different ranks.

3.3 Fusion of the Ranks

Table 1 overviews the runs where ■ shows which method is selected to fuse. The methods leverage different metrics to retrieve or re-rank the documents. Therefore, we merge them by using reciprocal rank fusion, which ignores document scores produced the retrieval system and takes only ranks into account (Cormack et al., 2009). We use the TrecTools (Palotti et al., 2019) to run the fusion algorithm.

4 Evaluation

In this section, we discuss the TREC results. The organizers provided the results from the baseline model, which is BM25. Before proceeding with evaluation, it is worth mentioning that we could

No	Criteria
1	Does the story adequately discuss the costs of the intervention?
2	Does the story adequately quantify the benefits of the intervention?
7	Does the story compare the new approach with existing alternatives?
8	Does the story establish the availability of the treatment/test/product/procedure?

Table 2: Criteria from Health News Review (https://www.healthnewsreview.org/about-us/review-criteria/)

ID	Name	Metric
1	graded.usefulness	nDCG
2	binary.useful-correct	nDCG
3	binary.useful-correct*	P@10
4	binary.useful-credible	nDCG
5	useful-correct-credible	nDCG
6	2aspects.correct-credible	CAM_MAP
7	2aspects.useful-credible	CAM_MAP
8	3aspects	CAM_MAP_3

Table 3: Mapping credibility, usefulness, and correctness evaluation to the evaluation metrics. nDCG: Normalized Discounted Cumulated Gain, CAM: Convex Aggregating Measure

not execute our models on the whole corpora due to a lack of computational resources. We obtained the results from the subset of the corpora. Therefore, we compare the results with our baseline (upv_bm25), not the official baseline.

The TREC evaluated the submitted runs based on the multiple aspect measures (Lioma et al., 2017) and the compatibility (Clarke et al., 2020c,d).

Tables 3 and 4 present multiple aspect scores and binary relevance scores of the models. As expected, the models could not compete with the scores from the official baseline. When we compare the models with upv_bm25 with the other models, upv_fuse_2 which combines the ranks of BM25 and SBERT has slightly performed better than the upv_bm25. SBERT could take the words' context into account, and better estimate the similarity between sentences having domain-specific words. This improves the retrieval of useful documents. However, integrating the quality estimators has not shown a positive effect on the results, and hence, we got scores worse than our baseline. It seems likely that the four criteria might be ineffective in assessing the quality of the documents. Also, we noticed that according to TREC assessing guidelines[5], quality estimation could overlap only with a few criteria in the list. For example, criteria such as estimating the credibility of the web pages and the sources are other parameters for consideration. Therefore, future work should include these modalities into re-ranking models and support more criteria in the Health News Review guidelines.

The other evaluation is to measure the harmfulness and helpfulness compatibility of the models. The best model should have lower harmfulness and higher helpfulness compatibility score. According to data in Table 5, upv_fuse_2 is the model which has the highest helpful compatibility score. Also, integrating the $QE^{\diamond}$ with the BM25 search engine and semantic search engine (upv_fuse_5 and upv_fuse_9) could slightly reduce the harmful scores. It seems that $QE^{\diamond}$ could filter out a few low-quality documents, which is promising for future work. However, it also negatively influences helpfulness compatibility. As we mentioned in the previous paragraph, a potential solution is to encode more quality criteria and other aspects of credibility.

5 Related Work

This section overviews the related studies.

5.1 Quality Estimation of Health Articles

DISCERN (Charnock et al., 1999) and the Health News Review checklist are the popular schema to assess the quality of health articles. However, DISCERN is used only for articles related to disease treatments. Unlike DISCERN, the checklist of the Health News Review applies to diverse topics in the medical domain, and the related datasets are

[5] https://trec-health-misinfo.github.io/docs/TREC-2021-Health-Misinformation-Track-Assessi Version-2.pdf, accessed on 27.10.2021

Models	1	2	3	4	5	6	7	8
official baseline	**0.5815**	0.3088	**0.4279**	**0.4867**	**0.3813**	**0.1605**	**0.2357**	0.205
upv_bm25	0.5285	0.3441	0.3828	0.445	0.3321	0.1452	0.2107	0.184
upv_fuse_2	**0.5316**	0.3412	**0.3959**	0.4413	**0.3345**	0.1256	**0.2108**	**0.1858**
upv_fuse_3	0.5127	0.2794	0.3666	0.4322	0.3176	0.12	0.1875	0.162
upv_fuse_5	0.5038	0.2529	0.3584	0.4296	0.3112	0.1315	0.1795	0.1539
upv_fuse_7	0.5204	0.2941	**0.3835**	0.4338	0.3287	0.1315	0.1964	0.1712
upv_fuse_9	0.5185	0.2941	0.3743	0.4303	0.3173	0.1276	0.193	0.1675

Table 4: Credibility, usefulness, and correctness of the models regarding multiple aspects and binary relevance (The details of 1-8 are given in Table 3). Bold scores are better than our baseline (upv_bm25).

Models	Harmful	Helpful
official baseline	0.1445	0.1292
upv_bm25	0.1043	0.1341
upv_fuse_2	0.1084	**0.1378**
upv_fuse_3	0.1114	0.1093
upv_fuse_5	0.1061	0.1195
upv_fuse_7	0.1036	0.1286
upv_fuse_9	**0.1018**	0.109

Table 5: Harmfulness and helpfulness of the models. Bold scores are better than our baseline (upv_bm25).

publicly available for implementing a quality estimation classifier.

Researchers used the checklist to automize the quality estimation. Afsana et al. (2020) are the first who investigated traditional machine learning algorithms, namely, SVM, Naive Bayes, Random Forest and Ensemble Vote classifier on Health News Review dataset. Furthermore, they designed various features such as Tf-idf, LIWC, POS tags, citations on the text. (Zuo et al., 2021; Al-Jefri et al., 2020) further examined the transformer models on the dataset and compared them with traditional models. Among the transformer models, RoBERTa has shown good results (Zuo et al., 2021). In our paper, we integrate the RoBERTa-based classifiers as a part of the re-ranking task.

5.2 TREC 2020: Health Misinformation Track

The previous Health Misinformation Track focused on retrieval of COVID-19 related documents from the Common Crawl corpus (Clarke et al., 2020b). We briefly describe the best-performing models. One approach leveraged T5 models (Raffel et al., 2020) to re-rank the documents according to their stance and relevance scores (Pradeep et al., 2020). The other model (Lima et al., 2020) estimates the credibility and misinformation score of the documents by employing classifiers and using the scores for re-ranking the documents. Although this study inspires our method to penalize misleading documents, we estimate quality using the schema designated explicitly for health documents.

6 Conclusion

In this paper, we have described our submissions to the TREC Health Misinformation Track 2021. First, we implemented a BM25 and a semantic search engine based on a domain-specific SBERT. Also, we used quality estimators based on RoBERTa models trained on the Health News Review dataset as a reranking model. Finally, we merged the ranks from the different components to get the final lists. The current study results show that integrating the rank lists from BM25 and the semantic search engine could improve the scores; however, the quality estimators were inefficient. As future work, we plan to study the integration of the different modalities for determining credibility and encoding the other aspects of the Health News Review guideline.

Acknowledgements

The work of Paolo Rosso was in the framework of the MISMIS-FAKEnHATE on MISinformation & MIScommunication in social media: FAKE news and HATE speech(PGC2018-096212-B-C31) and XAI-DisInfodemics on eXplainable AI for disinformation and conspiracy detection during infodemics (PLEC2021-007681) research projects funded by the Spanish Minitry of Science and Innovation as well as IBERIFIER, the Iberian Digital Media Research and Fact-Checking Hub funded by the Eu-

ropean Digital Media Observatory (2020-EU-IA-0252). Also, we would like to thank Ronak Pradeep for his help in providing the subset of the index corpus to us.

References

Fariha Afsana, Muhammad Ashad Kabir, Naeemul Hassan, and Manoranjan Paul. 2020. Automatically assessing quality of online health articles. *IEEE Journal of Biomedical and Health Informatics*, 25(2):591–601.

Majed Al-Jefri, Roger Evans, Joon Lee, and Pietro Ghezzi. 2020. Automatic identification of information quality metrics in health news stories. *Frontiers in public health*, 8:953.

D Charnock, S Shepperd, G Needham, and R Gann. 1999. Discern: an instrument for judging the quality of written consumer health information on treatment choices. *Journal of Epidemiology & Community Health*, 53(2):105–111.

Charles L. A. Clarke, Maria Maistro, Saira Rizvi, Mark D. Smucker, and Guido Zuccon. 2020a. Overview of the trec 2020 health misinformation track. In *TREC*.

Charles L. A. Clarke, Saira Rizvi, Mark D. Smucker, Maria Maistro, and Guido Zuccon. 2020b. Overview of the TREC 2020 health misinformation track. In *TREC*, volume 1266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Charles L. A. Clarke, Mark D. Smucker, and Alexandra Vtyurina. 2020c. Offline evaluation by maximum similarity to an ideal ranking. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management*, CIKM '20, page 225–234, New York, NY, USA. Association for Computing Machinery.

Charles L.A. Clarke, Alexandra Vtyurina, and Mark D. Smucker. 2020d. Offline evaluation without gain. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, ICTIR '20, page 185–192, New York, NY, USA. Association for Computing Machinery.

Gordon V Cormack, Charles LA 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*, pages 758–759.

Lisa Neal Gualtieri. 2009. The doctor as the second opinion and the internet as the first. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, pages 2489–2498.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Lucas Chaves Lima, Dustin Brandon Wright, Isabelle Augenstein, and Maria Maistro. 2020. University of copenhagen participation in TREC health misinformation track 2020. In *TREC*, volume 1266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *SIGIR*, pages 2356–2362. ACM.

Christina Lioma, Jakob Grue Simonsen, and Birger Larsen. 2017. Evaluation measures for relevance and credibility in ranked lists. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '17, page 91–98, New York, NY, USA. Association for Computing Machinery.

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 preprint arXiv:1907.11692*.

Joao Palotti, Harrisen Scells, and Guido Zuccon. 2019. Trectools: an open-source python library for information retrieval practitioners involved in trec-like campaigns. SIGIR'19. ACM.

Ronak Pradeep, Xueguang Ma, Xinyu Zhang, Hang Cui, Ruizhou Xu, Rodrigo Nogueira, and Jimmy Lin. 2020. H2oloo at TREC 2020: When all you got is a hammer... deep learning, health misinformation, and precision medicine. In *TREC*, volume 1266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP/IJCNLP (1)*, pages 3980–3990. Association for Computational Linguistics.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zeynep Akkalyoncu Yilmaz, Charles L. A. Clarke, and Jimmy Lin. 2020. A lightweight environment for learning experimental IR research practices. In *SIGIR*, pages 2113–2116. ACM.

Chaoyuan Zuo, Qi Zhang, and Ritwik Banerjee. 2021. An empirical assessment of the qualitative aspects of misinformation in health news. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 76–81.