Improving online health search via multi-dimensional information quality models based on deep learning

Boya Zhang[1], Nona Naderi[2,3], Rahul Mishra[1] and Douglas Teodoro[1,2,3]

[1] Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland
[2] Department of Information Science, University of Applied Sciences and Arts of Western Switzerland HES-SO, Geneva, Switzerland
[3] SIB Swiss Institute of Bioinformatics, Geneva, Switzerland
Email(s):
boya.zhang@unige.ch (B. Zhang); nona.naderi@hesge.ch (N. Naderi); rahul.mishra@unige.ch (R. Mishra); douglas.teodoro@unige.ch (D. Teodoro)
ORCID(s):
0000-0002-4439-8212 (B. Zhang); 0000-0002-1272-7640 (N. Naderi); 0000-0001-7511-2718 (R. Mishra); 0000-0001-6238-4503 (D. Teodoro)

Abstract

Background: The presence of widespread misinformation in Web resources and the limited quality control provided by search engines can lead to serious implications for individuals seeking health advice.
Objective: We aimed to investigate a multi-dimensional information quality assessment model based on deep learning to enhance the reliability of online healthcare information search results.
Methods: In this retrospective study, we simulated online health information search scenarios with a topic set of 35 different health-related inquiries and a corpus containing one billion Web documents from the April 2019 snapshot of Common Crawl. Using state-of-the-art pre-trained language models, we inferred the usefulness, supportiveness, and credibility quality dimensions of the retrieved documents for a given search query.
Results: The usefulness model provided the largest distinction between help and harm compatibility documents with a difference of 0.053. The supportiveness model achieved the best harm compatibility (0.024), while the combination of usefulness, supportiveness, and credibility models achieved the best help and harm compatibility on helpful topics.
Conclusions: Our results suggest that integrating automatic ranking models created for specific information quality dimensions can increase the effectiveness of health-related information retrieval for decision-making.

Keywords Health Misinformation; Information Retrieval; Deep Learning; Language Model; Transfer Learning.

1. Introduction

In contemporary society, people with varying information needs, medical knowledge, and linguistic skills [1] use the Web to seek health advice and make treatment decisions [2]. Due to the mixture of facts and rumours in online resources [3], Web users may have difficulty truly identifying the correct information and may thus be misled by false content [4]. The central issue in health information search is that Web search engines retrieve and rank documents mostly based on their relevance to the search query [5], ignoring information quality aspects. To provide high-quality resources and enable properly informed decision-making [6], information retrieval systems must be able to differentiate between accurate and inaccurate content [7].

Health misinformation is defined as health-related information that is inaccurate or misleading based on current scientific evidence [8,9]. Health misinformation has become increasingly prevalent online due to the lack of health literacy [10] of non-professionals and due to the infodemic phenomenon [11], that is, the rapid and wide spread of both accurate and inaccurate information about a medical topic in modern internet [12]. Recently, health misinformation-related topics, like "vaccine" or "the relationship between coronavirus and 5G", on social media platforms, such as Twitter and Instagram[13–15], and across countries [16] have gained scientific interest. Thus, new credibility-centred search methods and assessment measures are crucial for tackling open issues in health-related information retrieval [17]. In the past few years, information retrieval shared tasks, such as the Text REtrieval Conference (TREC) and the Conference and Labs of the Evaluation Forum (CLEF), have started evaluating quality-based systems for health corpora [18,19]. The CLEF eHealth Lab Series proposed a benchmark to evaluate models according to the relevance, readability, and credibility of the retrieved information [20]. The TREC Health Misinformation Track (2019-2021) proposed further metrics to evaluate models according to the relevance, efficacy, credibility, usefulness, correctness, and supportiveness of the retrieved information [21–23]. These metrics are one of the first efforts to quantitatively assess the effectiveness of information retrieval engines to retrieve high-quality information, moving away from the traditional query-document relevance paradigm [24,25].

Thanks to these open shared tasks, in the past years many prominent approaches have been developed to improve the search for higher-quality health information. In the absence of sufficiently large training data, classical bag-of-words-based methods tend to outperform neural network approaches in the detection of health-related misinformation [26]. On the other hand, a combination of classical handcrafted feature-based methods and deep learning-based methods can be effective when customized for health misinformation detection in Web content [17]. Multi-stage ranking systems [27,28] coupled with a label prediction model [29] improve upon bag-of-word-based baselines and have set the state-of-the-art [28] for separating helpful from harmful Web content [30].

In this study, we aim to investigate the impact of usefulness, supportiveness, and credibility dimensions on improving the quality of the retrieved health-related information. We propose an unsupervised multi-dimensional ranking model by utilizing deep learning-based pre-trained language models [31] through transfer learning, which categorizes the quality of the retrieved Web resources by adhering to different information quality dimensions. Specialized quality-oriented ranks obtained by re-ranking components are fused [32] to provide the final ranked list. In contrast to prior studies, our approach relies on the automatic detection of harmful (or inaccurate) claims and uses a multi-dimensional information quality model to boost helpful resources. Particularly, we study the ability of the proposed models to reinforce helpful documents to and deter harmful ones from the top ranks for health-related topics and show the impact of each information dimension on the rank quality.

2. Methods

2.1 Dataset

To evaluate our models, we use the TREC Health Misinformation Track 2021 benchmark [22] organized by the National Institute of Standards and Technology (NIST), which simulates Web searches for specific health issues and interventions, and further annotates the results according to their quality. The topics are searched against a collection of English Web documents [33]. For each topic, the benchmark annotates the retrieved Web documents using a pooling approach, where the top retrieved documents are evaluated according to their usefulness, correctness and credibility, and subsequently labelled as helpful or harmful[23].

2.1.1 Health-related Topics

A topic consists of a health issue, an intervention, a query that connects the corresponding intervention to the health problem, and a description that resembles the Web search question using natural language. As indicated in Table 1, there are a total of 35 topics in the benchmark, 17 of which are marked with the stance as helpful and 18 as unhelpful. The term 'helpful topic' refers to an intervention that is beneficial to the treatment of a health issue, while 'unhelpful topic' indicates that the intervention is ineffective to treat the health problem. The stance is supported by evidence from a reliable scientific source.

2.1.2 Web Corpus

In our study, we use the C4 corpus, which is a collection of English-language Web documents from the public Common Crawl web scrape [34]. We use the C4 English noclean version, which is 2.3TB in size and comprises one billion English documents from the April 2019 snapshot.

To give an example of the corpus content and how Web information can be contradicting, in Table 2 we present two documents relevant to the topic *'tepid sponge bath reduce fever in children'*. In this example, according to an article, the suggestion

is not to use the intervention (*'Do Not Use Sponging to Reduce a Fever'*) while it could be an option (*'Sponging is an option for high fevers'*) according to another.

Table 1. NIST helpful and unhelpful topics with query and description.

#	Query	Description
	helpful topics (N = 17)	
106	vitamin b12 sun exposure vitiligo	Can vitamin b12 and sun exposure together help treat vitiligo?
107	yoga asthma	Does yoga improve the management of asthma?
115	magnesium migraine prevention	Can magnesium help prevent migraine?
118	dupixent eczema	Can dupixent treat eczema?
120	imitrex migraine	Can the drug Imitrex (sumatriptan) treat acute migraine attacks?
121	light therapy lamp depression	Will buying a light therapy lamp help treat depression?
127	aromatherapy massage rheumatoid arthritis	Can aromatherapy massage help manage rheumatoid arthritis?
129	minoxidil balding hair growth	Can minoxidil treat hair loss?
131	l-theanine supplements anxiety	Can l-theanine supplements reduce stress and anxiety?
133	exercise depression	Does exercise improve the symptoms of depression?
136	dates iron deficiency anemia	Can eating dates help manage iron deficiency anemia?
140	fungal cream athlete's foot	Can fungal creams treat athlete's foot?
143	tylenol osteoarthritis	Does Tylenol manage the symptoms of osteoarthritis?
144	music therapy depression	Can music therapy help manage depression?
145	omega 3 borderline personality disorder	Can omega-3 treat borderline personality disorder in women?
146	vitamin d asthma attacks	Can vitamin D supplements improve the management of asthma?
149	hip osteoarthritis at-home exercises	Will at-home exercises manage hip osteoarthritis pain?
	unhelpful topics (N = 18)	
101	ankle brace achilles tendonitis	Will wearing an ankle brace help heal achilles tendonitis?
102	tepid sponge bath reduce fever children	Is a tepid sponge bath a good way to reduce fever in children?
103	folic acid dementia	Can folic acid help improve cognition and treat dementia?
104	duct tape warts	Does duct tape work for wart removal?
105	put ice on a burn	Should I apply ice to a burn?
108	starve a fever, feed a cold	Is starving a fever effective?
109	selenium cancer	Does selenium help prevent cancer?
110	birth control pill ovarian cysts treatment	Will taking birth control pills treat an ovarian cyst?
111	zinc supplements pregnancy	Will taking zinc supplements improve pregnancy?
112	evening primrose oil eczema	Can evening primrose oil help treat eczema?
114	vitamin e cream for skin scars	Can applying vitamin E cream remove skin scars?
117	fermented milk blood pressure	Can fermented milk help mitigate high blood pressure?
122	aleve migraine	Does Aleve relieve migraine headaches?
128	steam shower croup	Does steam from a shower help croup?
132	inhaling steam common cold	Does inhaling steam help treat common cold?
134	remove tick with vaseline	Can I remove a tick by covering it with Vaseline?
137	vinegar fish bone stuck	Will drinking vinegar dissolve a stuck fish bone?
139	copper bracelets reduce pain	Can copper bracelets reduce the pain of arthritis?

2.1.3 Query Relevance

As shown in Table 3, topic-Web document pairs were assessed by human annotators using three criterion dimensions - usefulness, correctness, and credibility – with different granularity levels. The usefulness (or relevance) of a document is examined first to decide whether it should be included in the query relevance set. Then, if a document is judged as correct (rating score between 7 and 12) or neutral (rating score between 1 and 6), it means that it is helpful for the treatment decision. Analogously, if the document is judged as incorrect (rating score between -1 and -3), then it would be harmful to the treatment decision. Further details about query relevance construction can be found in the annotation guidelines [35].

The distribution of helpful and harmful documents for each topic in the evaluation set as well as their respective average assessment scores are shown in Figure 1. For helpful topics, the number of helpful and harmful documents is not balanced, with a superior number of helpful documents (97% helpful). For unhelpful topics, the number of helpful and harmful documents is better balanced with helpful ones slightly outnumbering (97% vs. 54% helpful). The average assessment score for helpful documents is 6 while the average assessment score for harmful documents is -2.

Table 2. Example of contradictory useful documents for Topic 102 'Is a tepid sponge bath a good way to reduce fever in children?'.

Segment	Content	
Doc ID	en.noclean.c4-train.07165-of-07168.96468	en.noclean.c4-train.00001-of-07168.126948
Timestamp	2019-04-25T18:00:17Z	2019-04-23T20:13:31Z
Text	[...] Do Not Use Sponging to Reduce a Fever. It is not recommended that you use sponging to reduce your child's fever. There is no information that shows that sponging or tepid baths improve your child's discomfort associated with a fever or an illness. Cool or cold water can cause shivering and increase your child's temperature. Also, never add rubbing alcohol to the water. Rubbing alcohol can be absorbed into the skin or inhaled, causing serious problems such as a coma. [...]	[...] Sponging With Lukewarm Water: Note: Sponging is an option for high fevers, but not required. It is rarely needed. When to Use: Fever above 104° F (40° C) AND doesn't come down with fever meds. Always give the fever medicine at least an hour to work before sponging. How to Sponge: Use lukewarm water (85 - 90° F) (29.4 - 32.2° C). Sponge for 20-30 minutes. If your child shivers or becomes cold, stop sponging. [...]
URL	https://patiented.solutions.aap.org/	https://childrensclinicofraceland.com/

Table 3. Dimensions for assessing topic-Web document pairs and their respective assessment score [23].

Correctness	Correct		Neutral		Incorrect
Usefulness / Credibility	Very useful	Useful	Very useful	Useful	Very useful, Useful
Excellent	12	11	6	5	-3
Good	10	9	4	3	-2
Low	8	7	2	1	-1
Document	Helpful				Harmful

2.2 Quality-based Multi-Dimensional Ranking Model

The quality-based multi-dimensional ranking model proposed in this work is presented in Figure 2. The information retrieval process can be divided into two phases: retrieval and re-ranking. In the retrieval phase, for a given topic j, Web documents are retrieved using a BM25 model [36]. The number of documents N_D per topic was set to 10,000 in our experiments. Then, in the re-ranking phase, we further estimate the relevance and quality of the retrieved documents according to the

usefulness, supportiveness, and credibility dimensions. To deal with the lack of training data, multiple unsupervised models are used for each dimension and their results are combined using a ranking fusion method for each ranking module.

Figure 1. Distribution of helpful/unhelpful topics according to helpful/harmful documents in the query relevance set.

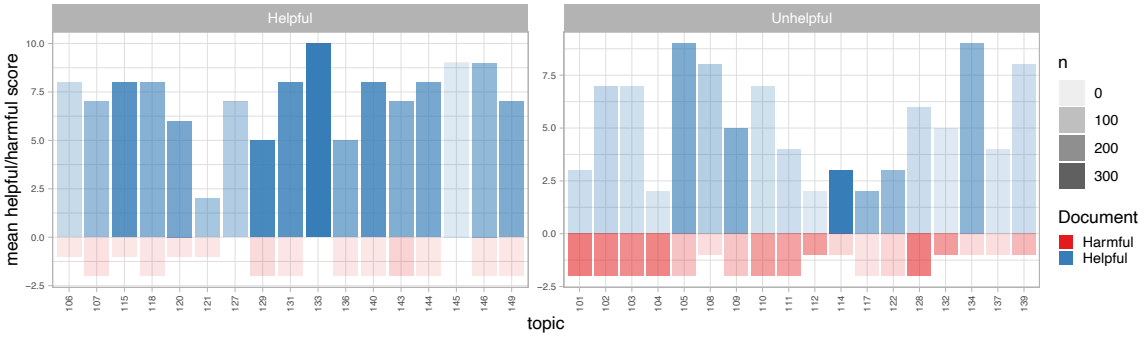
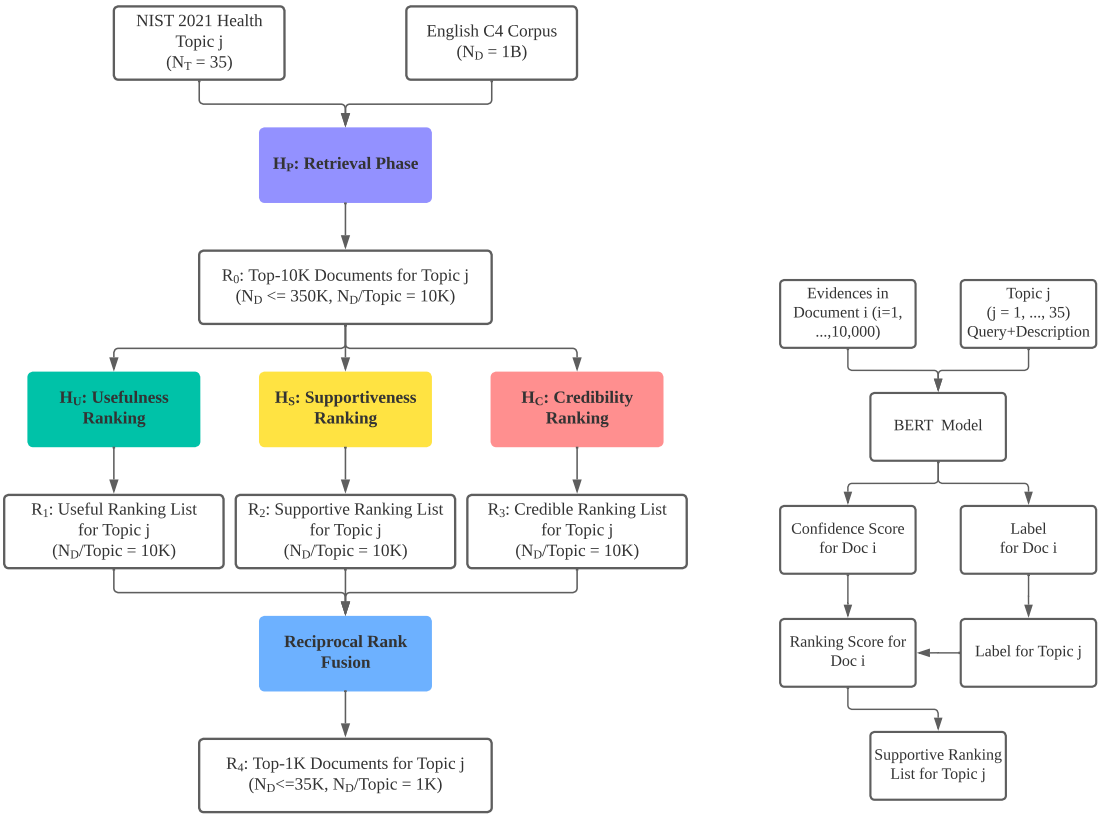


Figure 2. Quality-based multi-dimensional ranking models.



(a) General Pipeline (b) Supportiveness Model

2.2.1 Retrieval Phase

Two indices are created for the C4 snapshot collection, one using standard BM25 [36] parameters and another is fine-tuned using a collection of topics automatically

generated (silver standard) from a set of 4,985 indexed documents. For a given document, the silver topic was created based on the keyword2query [37] and doc2query [38] models to provide the query and description content, respectively. Then, using the silver topics and their respective documents, the BM25 parameters of the second index are fine-tuned using grid search in a known-item search approach [39], i.e., for a given silver topic, the model should return on top-1 the respective document used to generate it. Finally, for a given topic of the benchmark collection, 10,000 documents are retrieved from both standard BM25 index and fine-tuned indices.

2.2.2 Re-Ranking Phase

To provide higher quality documents at the top ranks, we use a set of machine learning models trained to classify documents according to the usefulness, supportiveness, and credibility dimensions. For the initial lists obtained in the retrieval phase, the documents were re-ranked according to the following strategies:

Usefulness: The usefulness dimension defines how pertinent a document is to a given topic. To compute the usefulness of retrieved documents, we implement topic-document similarity models based on pre-trained language models such as BERT-base [40], mono-BERT-large [38], and ELECTRA [41], which are fine-tuned with the MS MARCO dataset. Given a topic-document pair, the language model infers a score that gives the level of similarity between the two input text passages. While BM25 provides a strong baseline for usefulness, it does not consider word relations by learning context-sensitive representations as is the case with the pre-trained language models, which are used to enhance the quality of the original ranking [25].

Supportiveness: The supportiveness dimension defines how accurate is the information provided by the document. In this dimension, documents are identified under three levels: 1) *supportive*, i.e., the document supports the treatment; 2) *dissuasive*, i.e., the document refutes the treatment; and 3) *neutral*, i.e., the document does not contain enough information to make the decision [22]. We want documents that are either supportive (if the topic is helpful) or dissuasive (if the topic is unhelpful) on the top of the ranking list, which means that correct documents are boosted, and misinforming documents are downgraded. Claim-checking models [42], which take a claim and a document as the information source and validate the veracity of the claim based on document content are applicable to the supportiveness dimension. Most claim-checking models assume that document content is ground truth. Since this is not valid in the case of Web documents, we add a further classification step, which evaluates the correctness of the retrieved documents. We use the top-k assignments [43] provided by the claim-checking models (k is set to 10,000 in our experiments) to define whether the topic should be supported or refuted. The underlying assumption is that a scientific fact is defined by the largest number of evidence available for a topic. Then, a higher rank is given to the correct supportive/dissuasive documents, a medium rank is given to the neutral documents and a lower rank is given to the incorrect supportive/dissuasive documents. A pictorial description of the process is provided in Figure 2b. We use three models

from the Scientific Claim Verification task [44] to classify the topics: RoBERTa-Large [45], BioMedRoBERTa-base [46] and SciBERT-base [47]. These models are trained on either scientific or large English corpora and fine-tuned on the FEVER [48] and SciFact [44] datasets.

Credibility: The credibility dimension defines how trustworthy is the source document. For estimating this dimension, we develop a random forest classifier trained on the Microsoft Credibility dataset [49] with a set of credibility-related features, such as readability, openpage rank [50] and number of cascading style sheets (CSS). The dataset consists of 1,000 Web pages on five topics of Health, Politics, Finance, Environmental Science, and Celebrity News. The Web pages are manually rated with credibility scores between 1 ("very non-credible") and 5 ("very credible").[1] We convert these scores for a binary classification setting – that is, the scores of 4 and 5 are considered as 1 or *credible* and scores of 1, 2, and 3 are considered as 0 or *non-credible*. For the readability score, we rely on the Simple Measure of Gobbledygook (SMOG) index [51], which estimates the years of education an average person needs to understand a piece of writing. Following Schwarz and Morris [49], we retrieve a Web page's PageRank and use it as a feature to train the classifier. We further use the number of CSS style definitions for estimating the effort for the design of a Web page (Olteanu *et al*. [52] showed the effectiveness of this feature). Furthermore, a list of credible websites scrapped from the Health On the Net search engine [53] for the evaluated topics is combined with the baseline model to explore better performance. The result of the classifier is added to the unitary value of the Health On the Net credible sites [54].

2.3 Evaluation Method

We employ the "compatibility" metric [55] to assess the performance of our models. Contrary to the classic information retrieval tasks, in which the performance metric relies on the degree of relatedness between queries and documents, in quality retrieval, harmful documents should be penalised, especially if they are relevant to the query content. The compatibility metric calculates the similarity between the actual ranking R provided by a model and an ideal ranking I as provided by the query relevance annotations. According to Equation 1, the compatibility is calculated with the Rank Biased Overlap (RBO) [56] similarity metric, which is top-weighted, with greater weight placed at higher ranks to address the indeterminate and incomplete nature of Web search results [57]:

$$RBO(R, I) = (1 - p) \sum_{i=1}^{K} p^{i-1} \frac{|I_{1:i} \cap R_{1:i}|}{i} \quad (1)$$

where the parameter p represents the searcher patience or persistence and is set to 0.95 in our experiments, and K is the search depth and is set to 1,000 to bring $pK-1$ as close to 0 as possible. As shown in Equation 2, an additional normalization step is

[1] A credible Web page is defined as "a page whose information one can accept as the truth without needing to look elsewhere.

added to accommodate short, truncated ideal results, so when there are fewer documents in the ideal ranking than in the actual ranking list, it does not influence the compatibility computation results:

$$NRBO(R,I) = \frac{RBO(R,I)}{RBO(I,I)} \tag{2}$$

To ensure that helpful and harmful documents are treated differently, even if both might be relevant to the query content, the assessments are divided into "help compatibility" (help) and "harm compatibility" (harm) metrics. Then, to evaluate the ability of the system to separate helpful from harmful information, the "harm compatibility" results are subtracted from the "help compatibility" results, which are marked as "help-harm compatibility" (help-harm). Overall, the more a ranking is compatible with the ideal helpful ranking, the better it is. Conversely, the more a ranking is compatible with the ideal harmful ranking, the worse it is.

Table 4. Help, harm, and help-harm compatibility for the quality-based retrieval models stratified by helpful, unhelpful, and all topics. H_U: Usefulness; H_S: Supportiveness; H_C: Credibility.

Model	Help ↑			Harm ↓			Help-Harm ↑		
	$helpful_T$	$unhelpful_T$	all_T	$helpful_T$	$unhelpful_T$	all_T	$helpful_T$	$unhelpful_T$	all_T
BM25 [36]	-	-	0.122	-	-	0.144	-	-	-0.022
Pradeep *et al.* [29]	-	-	**0.195**	-	-	0.153	-	-	0.043
Abualsaud *et al.* [58]	-	-	0.164	-	-	0.123	-	-	0.040
Schlicht *et al.* [59]	-	-	0.121	-	-	0.103	-	-	0.018
Pichel *et al.* [60]	-	-	0.163	-	-	0.155	-	-	0.008
Bondarenko *et al.* [61]	-	-	0.129	-	-	0.144	-	-	-0.015
H_U	0.190	**0.093**	0.140	0.050	**0.115**	**0.087**	0.140	**-0.022**	**0.053**
$H_U + H_S$	0.184	0.077	0.129	0.048	0.126	0.092	0.136	-0.049	0.037
$H_U + H_S + H_C$	**0.196**	0.078	0.136	**0.043**	0.135	0.095	**0.153**	-0.057	0.041

Table 5. Help, harm, and help-harm compatibility for the individual quality-based models stratified by *helpful*, *unhelpful*, and *all* topics. H_p': Retrieval; H_U': Usefulness; H_S': Supportiveness; H_C': Credibility. H_U', H_S' and H_C' rankings are not combined with H_p as opposed to H_U, H_S and H_C.

Model	Help ↑			Harm ↓			Help-Harm ↑		
	$helpful_T$	$unhelpful_T$	all_T	$helpful_T$	$unhelpful_T$	all_T	$helpful_T$	$unhelpful_T$	all_T
H_p	**0.176**	**0.106**	**0.140**	0.026	0.178	0.111	**0.150**	-0.072	0.029
H_U'	0.158	0.065	0.110	0.074	0.084	0.080	0.084	-0.019	**0.030**
H_S'	0.047	0.015	0.031	0.018	**0.028**	**0.024**	0.029	**-0.013**	0.007
H_C'	0.042	0.017	0.029	**0.017**	0.048	0.035	0.025	-0.031	-0.006

3. Results

The help, harm, and help-harm compatibility for different combinations of quality-based retrieval models are shown in Table 4. The proposed approach provides state-of-the-art results for unsupervised automatic ranking systems, with help-harm compatibility of 0.053. The result is obtained by the usefulness model (H_U), which as shown in Figure 2 is the combination of retrieval and re-ranking steps, the latter using

a pre-trained neural language model fine-tuned for the usefulness dimension. Among our models, the usefulness model also obtains the best help and harm compatibility (0.140 and 0.087, respectively). Notice that for the latter metric, the closest to 0, the best is the performance. Interestingly, the combination of usefulness (H_U), supportiveness (H_S), and credibility (H_C) models provides the best help (0.196) and harm (0.043) compatibility for helpful topics. Finally, our models outperform a BM25-based baseline for the three compatibility metrics and also provide state-of-the-art results for the harm compatibility metric, showing their robustness for refuting harmful documents, while providing relatively good performance for helpful documents in Web searches for health-related content.

3.1 Performance Stratification by Quality Dimension

Table 5 shows the results of the quality-based models disregarding the retrieval step (prime index). The usefulness model (H_U') achieves the highest help-harm compatibility (0.30), close to the compatibility of the basic retrieval model (H_P = 0.029). We observe that the retrieval model gives the best result in terms of help compatibility (0.140 for all topics); however, it also shows the highest harm compatibility (0.111 for all topics). Thus, it is slightly less capable of distinguishing between correct and incorrect information as compared to the usefulness model alone. Their combination (i.e., H_U in Table 4) improves the retrieval model by a remarkable 83% (from 0.029 to 0.053). For the harm compatibility, the supportiveness model achieves the overall best performance among the individual models (0.024) but also for unhelpful topics (0.028). In addition, with a score of -0.013, the model is best on the help-harm compatibility for unhelpful topics, which demonstrates the ability of the supportiveness model to better differentiate between correct and incorrect information for unhelpful topics. Lastly, with a score of 0.017, the credibility model outperforms the other models on harmful compatibility for unhelpful topics. Overall, the combination of the usefulness, supportiveness, and credibility models improves the help-harm compatibility with both the helpful and harmful topics (2% and 20% relative improvements, respectively) as compared to the retrieval step, leading thus to better identification of correct information for helpful topics, incorrect information for harmful topics, and both correct and incorrect information for all topics.

3.2 Quality Control

One of the advantages of the proposed multi-dimensional model as compared to other approaches in the literature is that we can optimize the results according to different quality metrics. In Figure 3, we show how the compatibility score varies with the weight changes of specific models. The final result is the combination of four models targeting three different dimensions. This includes the H_P, H_U', H_S', and H_C'. We vary the weight of one specific model from 0 to 2 and fix three other models with a weight of 0.33 to balance, and eventually, we see how the effect of each individual model is added to the final result. When altering the weight of the retrieval model H_P as indicated by the green line, both the helpful and harmful compatibility have an obvious increase before the weight is added to 0.5 and the increasing trend became smooth from 0.5 onwards to 2. For the help-harm compatibility score, we also achieve

the highest value at a weight of 0.5 with the value slightly decreasing after surpassing the weight at 1.

When altering the weight of the individual usefulness model H_U', the helpful compatibility increases until the weight is 0.5 while the harmful compatibility increases until the weight is 0.4. From those points on, both the helpful and harmful compatibility decrease with the helpful-harmful compatibility increasing until the weight is 0.8 and the result of adding this weight achieves the same performance as H_U, which is the combination of retrieval and usefulness model.

For both the H_S' and H_C', the compatibility decreases with the weight addition regardless of whether it is helpful compatibility, harmful compatibility, or the difference between the two. We would expect the models to decrease the harmful compatibility while increasing helpful compatibility. So far on a general basis, our model only could increase the helpful compatibility more than the harmful compatibility or decrease the harmful compatibility more than the helpful compatibility.

Figure 3. Help, harm, and help-harm compatibility with weights added to specific model H_P, H_U', H_S' and H_C'

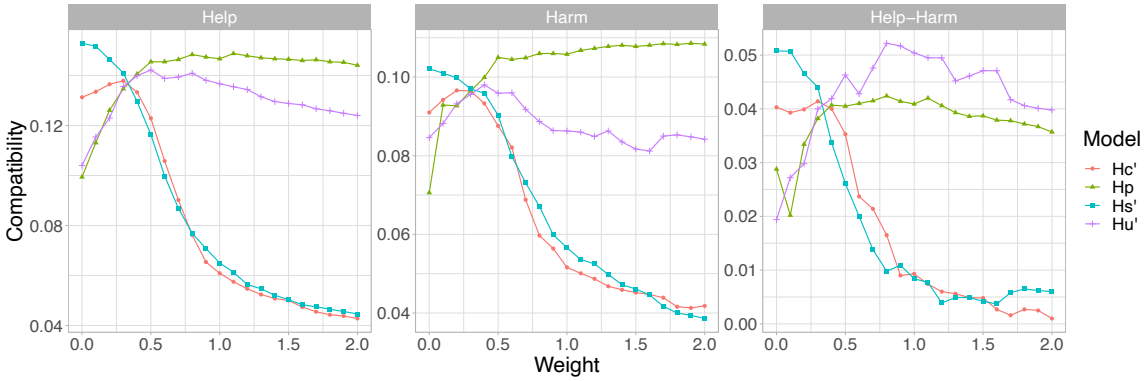
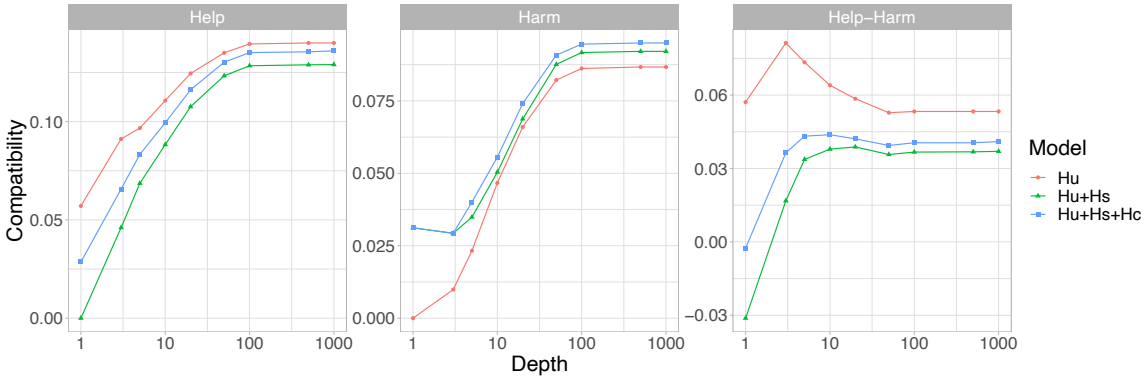


Table 6. Help, harm, and help-harm compatibility with search depth set to 10.

Model	Help ↑	Harm ↓	Help-Harm ↑
H_U	0.111	0.047	0.064
$H_U + H_S$	0.088	0.050	0.038
$H_U + H_S + H_C$	0.099	0.056	0.044

Figure 4. Help, harm, and help-harm compatibility with different search depth.



3.3 Model Interpretation

To semantically explain the variation of help-harm compatibility, we set the search depth K to 10 in Equation 1. The help, harm, and help-harm compatibility of the three models are shown in Table 6. The help-harm compatibility is 1 when only helpful documents are retrieved in the top 10. Conversely, the help-harm compatibility is -1 when only harmful documents are retrieved in the top 10. A variation of 0.1 in the help or harm compatibility corresponds roughly to one helpful document exceeding the number harmful documents retrieved in the top 10. Overall, the results show that retrieving relevant documents for the queries is hard as on average only 1.5 out of 10 documents are relevant (helpful or harmful) to the topic. In addition, we interpret that the three models retrieve on average twice the number of helpful documents as compared harmful documents. Particularly, H_U has on average around one more helpful than harmful document in the top 10.

In Figure 4, we show the average compatibility for all the topics as the search depth K varies. When the search depth is 1, we compare the top-1 document in the ranking list and the query relevance set. In this case, the compatibility is either 0 or 1 for each topic. When the search depth is above 1, we compare the top-K documents and the compatibility for each topic varies between 0 and 1. The compatibilities increase almost monotonically until the search depth K becomes 100 as the query reaches the average number of documents in the relevance set (i.e., 92 documents) and subsequently becomes saturated. Interestingly, we observe that for help-harm, the model H_U spikes at the depth of 3 due to the higher proportion of helpful documents retrieved as compared to harmful documents. Further, with the depth increase, we notice that the difference in proportions of helpful and harmful documents diminishes steeply between depths 3 to 10.

4. Discussion

Information retrieval systems without quality measurement paint an incomplete picture of Web search for health information. In this paper, we propose a quality-based multi-dimensional ranking model by utilizing pre-trained language models to enhance the reliability of retrieved documents according to usefulness, supportiveness, and credibility dimensions. By combining different re-ranking strategies, we show that the help compatibility increases more than the harm compatibility on helpful topics and the harm compatibility decreases by a substantial

amount compared to help compatibility on unhelpful topics. Consequently, we observe a 40% increase in the overall help-harm compatibility and state-of-the-art performance for automatic ranking systems.

Comparing the compatibilities among default, automatic, and manual retrieval strategies proposed in the literature, results show that the default BM25 model [36] model with -0.022 help-harm compatibility is the lower bound for automatic models on all topics with a mixture of helpful and harmful documents when we put no effort in detecting health misinformation. On the other hand, the help-harm compatibility for the three best automatic models found in the literature [29,31,58] range from 0.040 to 0.043. Automatic models prioritize helpful documents, resulting thus in positive help-harm compatibility. These results are far from state-of-art manual models, with help-harm compatibility scores ranging from 0.208 to 0.259. The best manual model [29] re-ranks BM25 results using T5 [33] knowing the stance and the coincide reformulated description of the topics. The human model [58] using two assessors for interactive search with a continuous active learning method is a benchmark of human performance. We acknowledge that the help-harm compatibility can improve significantly with the correct stance given. However, this information is hardly available in standard search environments. Therefore, finding the correct stance automatically is another key component of the automatic model.

Previous research on health misinformation has given more emphasis to the detection of misinformation rather than improving the reliability of health information search results. These works typically utilize either handcrafted feature-based models [62] to identify the misinformation or use the deep learning-based approach in combination with some auxiliary information such as knowledge bases [63]. Other works improve the reliability of search results based on the importance and relevance of the documents to the topic or by adding dimensional metrics [58–61]. To retrieve accurate and reliable documents, we take three dimensions of usefulness, supportiveness, and credibility into account and present more comprehensive results. Pradeep *et al*. [29] use pre-trained language models for both relevance ranking and label prediction sequentially. Conversely, we execute these steps in parallel so that the ranking in each dimension can be adjusted independently of the others and can effectively improve the model's ability to distinguish between helpful and harmful documents.

In the query relevance set, helpful topics are often associated with more helpful as compared to harmful documents, while unhelpful topics are evenly associated with helpful and harmful documents. This indicates that relevant Web documents are biased in terms of quality. Web documents that discuss helpful topics tend to have a higher chance to be supportive of the intervention and are helpful while Web documents that discuss unhelpful topics tend to have an equal chance of being supportive or dissuasive on the intervention and are harmful or helpful, respectively. We observe that for all models, helpful topics achieve higher help compatibility, whereas unhelpful topics achieve higher harm compatibility. The average help-harm compatibility ranges from 0.136 to 0.153 for helpful topics and -0.057 to -0.022 for

unhelpful topics. Thus, the bias of helpfulness on retrieved documents for helpful topics is greater than the bias of harmfulness on retrieved documents for unhelpful topics. As a result, automatic retrieval systems tend to find more helpful information on helpful topics with the information biased towards helpfulness and find more harmful information on unhelpful topics with the information slightly biased towards harmfulness. Therefore, for the improvement of retrieval systems, it is important to focus on the quality aspects, especially for unhelpful topics.

This work has some limitations. First, in the supportiveness dimension, we use pre-trained language models to classify the concatenation of the evidence sentence and query-description pair into *Support*, *Refute,* and *Neutral*. More *Neutral* labels are obtained since the query-description pairs we input to the model are not in a claim style. We can reformulate descriptions from the question format to the claim format, evaluate the veracity of the claim, and determine the stance of the topic. Second, we make an improvement on the help-harm compatibility with a list of credible sites for the rankings on the credibility dimension. An alternative approach would be to put credibility before the supportiveness dimension in the re-ranking pipeline and use the retrieved documents only from credible sites to determine the stance of topics. However, the model may not be generalizable as the list of credible sites is hardly exhaustive.

Finally, the stance can be predicted with a credible corpus, but it can be relatively small and limited. We can use retrieved documents on the Web to determine the stance of the topics as helpful topics are often associated with more positive as compared to negative documents, while unhelpful topics are evenly associated with positive and negative documents.

To conclude, the proliferation of health misinformation in Web resources has led to mistrust and confusion among online health advice seekers. Automatic maintenance of factual discretion in Web search results is the need of the hour. In this work, we propose a multi-dimensional information quality assessment model which utilizes three information quality dimensions, namely usefulness, supportiveness, and credibility, to strengthen the factual reliability of the health advice search results. Experiments conducted on publicly available datasets show that the proposed quality assessment-based model is promising, performing well in comparison to various baselines and achieving state-of-the-art performance for automatic ranking. Thus, it could be used to improve online health search and to provide quality enhanced information for health information seekers.

Acknowledgements

Author contributions statement

BZ, NN, and DT prepared the data, conceived, and conducted the experiments, as well as analysed the results. BZ, NN, RM, and DT drafted the manuscript. All authors reviewed the manuscript.

Conflicts of Interest

The authors declare no competing interests.

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BM25: Okapi Best Match 25
CSS: cascading style sheets
CLEF: Conference and Labs of the Evaluation Forum
C4: Colossal, Cleaned Crawl Corpus
NIST: National Institute of Standards and Technology
NLP: natural language processing
RBO: Rank Biased Overlap
SMOG: Simple Measure of Gobbledygook
TREC : Text REtrieval Conference

References

1. Goeuriot Lorraine and Jones GJF and KL and MH and ZJ. Medical information retrieval: introduction to the special issue. Information Retrieval Journal 2016 Apr;19(1):1–5. [doi: 10.1007/s10791-015-9277-8]
2. Chu Joanna TW and Wang MP and SC and VK and LTH and CSSC. How, When and Why People Seek Health Information Online: Qualitative Study in Hong Kong. Interact J Med Res 2017 Dec;6(2):e24. [doi: 10.2196/ijmr.7000]
3. Lee Jung Jae and Kang K-A and WMP and ZSZ and WJYH and OS and YSC and SS. Associations Between COVID-19 Misinformation Exposure and Belief With COVID-19 Knowledge and Preventive Behaviors: Cross-Sectional Online Study. J Med Internet Res 2020 Nov;22(11):e22205. [doi: 10.2196/22205]
4. Ecker Ullrich K. H. and Lewandowsky S and CJ and SP and FLK and BN and KP and VEK and AMA. The psychological drivers of misinformation belief and its resistance to correction. Nature Reviews Psychology 2022 Jan;1(1):13–29. [doi: 10.1038/s44159-021-00006-y]
5. Sundin O, Lewandowski D, Haider J. Whose relevance? Web search engines as multisided relevance machines. J Assoc Inf Sci Technol 2022;73(5):637–642. [doi: https://doi.org/10.1002/asi.24570]
6. Krist Alex H. and Tong ST and ARA and LDR. Engaging Patients in Decision-Making and Behavior Change to Promote Prevention. Stud Health Technol Inform 2017;240:284–302.
7. Swire-Thompson B, Lazer D. Public Health and Online Misinformation: Challenges and Recommendations. Annu Rev Public Health 2020;41(1):433–451. [doi: 10.1146/annurev-publhealth-040119-094127]
8. di Sotto Stefano and Viviani M. Health Misinformation Detection in the Social Web: An Overview and a Data Science Approach. Int J Environ Res Public Health MDPI; 2022 Feb;19(4):2173. [doi: 10.3390/ijerph19042173]
9. Sylvia Chou W-Y, Gaysynsky A, Cappella JN. Where We Go From Here: Health Misinformation on Social Media. Am J Public Health 2020;110(S3):S273–S275. [doi: 10.2105/AJPH.2020.305905]

10. Kickbusch IS. Health literacy: addressing the health and education divide. Health Promot Int 2001 Sep;16(3):289–297. [doi: 10.1093/heapro/16.3.289]
11. Suarez-Lledo Victor and Alvarez-Galvez J. Prevalence of Health Misinformation on Social Media: Systematic Review. J Med Internet Res 2021 Jan;23(1):e17187. [doi: 10.2196/17187]
12. Eysenbach G. How to Fight an Infodemic: The Four Pillars of Infodemic Management. J Med Internet Res 2020 Jun;22(6):e21820. [doi: 10.2196/21820]
13. Burki T. Vaccine misinformation and social media. Lancet Digit Health 2019;1(6):e258–e259. [doi: https://doi.org/10.1016/S2589-7500(19)30136-0]
14. Lotto Matheus and Sá Menezes T and ZHI and TS-F and ABZ and PMP and CT. Characterization of False or Misleading Fluoride Content on Instagram: Infodemiology Study. J Med Internet Res 2022 May;24(5):e37519. [doi: 10.2196/37519]
15. Mackey TK, Purushothaman V, Haupt M, Nali MC, Li J. Application of unsupervised machine learning to identify and characterise hydroxychloroquine misinformation on Twitter. Lancet Digit Health 2021;3(2):e72–e75. [doi: https://doi.org/10.1016/S2589-7500(20)30318-6]
16. Nsoesie Elaine Okanyene and Cesare N and MM and O al. COVID-19 Misinformation Spread in Eight Countries: Exponential Growth Modeling Study. J Med Internet Res 2020 Dec;22(12):e24425. [doi: 10.2196/24425]
17. Upadhyay R, Pasi G, Viviani M. Health Misinformation Detection in Web Content: A Structural-, Content-Based, and Context-Aware Approach Based on Web2Vec. Proceedings of the Conference on Information Technology for Social Good New York, NY, USA: Association for Computing Machinery; 2021. p. 19–24. [doi: 10.1145/3462203.3475898]
18. Pogacar FA, Ghenai A, Smucker MD, Clarke CLA. The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments. Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval New York, NY, USA: Association for Computing Machinery; 2017. p. 209–216. [doi: 10.1145/3121050.3121074]
19. Upadhyay R, Pasi G, Viviani M. An Overview on Evaluation Labs and Open Issues in Health-related Credible Information Retrieval. 2021;
20. Suominen Hanna and Kelly L and GL and KM. CLEF eHealth Evaluation Lab 2020. In: Jose Joemon M. and Yilmaz E and MJ and CP and FN and SMJ and MF, editor. Advances in Information Retrieval Cham: Springer International Publishing; 2020. p. 587–594.
21. Abualsaud M, Lioma C, Maistro M, Smucker MD, Guido, Zuccon. Overview of the TREC 2019 Decision Track. 2020.
22. Clarke CLA, Rizvi S, Smucker MD, Maistro M, Zuccon G. Overview of the TREC 2020 health misinformation track. TREC 2020.
23. Clarke CLA, Maistro M, Smucker MD. Overview of the TREC 2021 Health Misinformation Track. TREC 2021.
24. Zhang E, Gupta N, Tang R, Han X, Pradeep R, Lu K, Zhang Y, Nogueira R, Cho K, Fang H, Lin J. Covidex: Neural Ranking Models and Keyword Search

Infrastructure for the COVID-19 Open Research Dataset. Proceedings of the First Workshop on Scholarly Document Processing [Internet] Online: Association for Computational Linguistics; 2020. p. 31–41. [doi: 10.18653/v1/2020.sdp-1.5]
25. Teodoro D, Ferdowsi S, Borissov N, Kashani E, Alvarez DV, Copara J, Gouareb R, Naderi N, Amini P, others. Information retrieval in an infodemic: the case of COVID-19 publications. J Med Internet Res JMIR Publications Inc., Toronto, Canada; 2021;23(9):e30161.
26. Fernández-Pichel Marcos and Losada DE and PJC and ED. Comparing Traditional and Neural Approaches for Detecting Health-Related Misinformation. In: Candan K. Selçuk and Ionescu B and GL and LB and MH and JA and MM and PF and FG and FN, editor. Experimental IR Meets Multilinguality, Multimodality, and Interaction Cham: Springer International Publishing; 2021. p. 78–90.
27. Lima LC, Wright DB, Augenstein I, Maistro M. University of copenhagen participation in trec health misinformation track 2020. arXiv preprint arXiv:210302462 2021;
28. Pradeep R, Ma X, Zhang X, Cui H, Xu R, Nogueira R, Lin J. H2oloo at TREC 2020: When All You Got Is a Hammer... Deep Learning, Health Misinformation, and Precision Medicine. Text Retrieval Conference (TREC) 2020.
29. Pradeep R, Ma X, Nogueira R, Lin J. Vera: Prediction Techniques for Reducing Harmful Misinformation in Consumer Health Search. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval New York, NY, USA: Association for Computing Machinery; 2021. p. 2066–2070. [doi: 10.1145/3404835.3463120]
30. Pradeep R, Ma X, Nogueira R, Lin J. Scientific Claim Verification with VerT5erini. Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis online: Association for Computational Linguistics; 2021. p. 94–103.
31. Zhang B, Naderi N, Jaume-Santero F, Teodoro D. DS4DH at TREC Health Misinformation 2021: Multi-Dimensional Ranking Models with Transfer Learning and Rank Fusion. 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.
32. Cormack G v, Clarke CLA, Buettcher S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval 2009. p. 758–759.
33. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research [Internet] 2020;21(140):1–67. Available from: http://jmlr.org/papers/v21/20-074.html
34. Common Crawl. [Internet]. Available from: https://commoncrawl.org/

35. TREC Health Misinformation Track (2022). [Internet]. 2018. Available from: https://trec-health-misinfo.github.io/
36. Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. Now Publishers Inc; 2009.
37. Bennani-Smires K, Musat C, Hossmann A, Baeriswyl M, Jaggi M. Simple Unsupervised Keyphrase Extraction using Sentence Embeddings. Proceedings of the 22nd Conference on Computational Natural Language Learning 2018. p. 221–229.
38. Nogueira R, Yang W, Cho K, Lin J. Multi-Stage Document Ranking with BERT. arXiv e-prints 2019;arXiv–1910.
39. Ogilvie P, Callan J. Combining Document Representations for Known-Item Search. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval [Internet] New York, NY, USA: Association for Computing Machinery; 2003. p. 143–150. [doi: 10.1145/860435.860463]
40. Li C, Yates A, MacAvaney S, He B, Sun Y. PARADE: Passage representation aggregation for document reranking. arXiv preprint arXiv:200809093 2020;
41. Clark K, Luong M-T, Le Q v, Manning CD. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:200310555 2020;
42. Stammbach D, Zhang B, Ash E. The Choice of Knowledge Base in Automated Claim Checking. ArXiv Cornell University; 2021;2111–7795.
43. Teodoro D, Gobeill J, Pasche E, Ruch P, Vishnyakova D, Lovis C. Automatic IPC encoding and novelty tracking for effective patent mining. The 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access 2010.
44. Wadden D, Lin S, Lo K, Wang LL, van Zuylen M, Cohan A, Hajishirzi H. Fact or Fiction: Verifying Scientific Claims. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020. p. 7534–7550.
45. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:190711692 2019;
46. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, Smith NA. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. Proceedings of ACL 2020.
47. Beltagy I, Lo K, Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) 2019. p. 3615–3620.
48. Thorne J, Vlachos A, Christodoulopoulos C, Mittal A. FEVER: a Large-scale Dataset for Fact Extraction and VERification. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) 2018. p. 809–819.

49. Schwarz J, Morris M. Augmenting web pages and search results to support credibility assessment. Proceedings of the SIGCHI conference on human factors in computing systems 2011. p. 1245–1254.
50. Open Page Rank API Documentation. [Internet]. 2022 [cited 2022 Sep 9]. Available from: https://www.domcop.com/openpagerank/documentation
51. Grabeel KL, Russomanno J, Oelschlegel S, Tester E, Heidel RE. Computerized versus hand-scored health literacy tools: a comparison of Simple Measure of Gobbledygook (SMOG) and Flesch-Kincaid in printed patient education materials. J Med Libr Assoc 2018 Jan;106(1):38–45.
52. Olteanu A, Peshterliev S, Liu X, Aberer K. Web credibility: Features exploration and credibility prediction. European conference on information retrieval 2013. p. 557–568.
53. Health On the Net, promotes transparent and reliable health information online through HONcode certification. [Internet]. [cited 2022 Sep 9]. Available from: https://myhon.ch/en/
54. Boyer C, Selby M, Scherrer J-R, Appel RD. The Health On the Net Code of Conduct for medical and health Websites. Comput Biol Med [Internet] 1998;28(5):603–610. [doi: https://doi.org/10.1016/S0010-4825(98)00037-7]
55. Clarke CLA, Vtyurina A, Smucker MD. Assessing Top-Preferences. ACM Transactions on Information Systems (TOIS) ACM New York, NY; 2021;39(3):1–21.
56. Webber W, Moffat A, Zobel J. A Similarity Measure for Indefinite Rankings. ACM Trans Inf Syst New York, NY, USA: Association for Computing Machinery; 2010 Nov;28(4). [doi: 10.1145/1852102.1852106]
57. Clarke CLA, Smucker MD, Vtyurina A. Offline Evaluation by Maximum Similarity to an Ideal Ranking. Proceedings of the 29th ACM International Conference on Information and Knowledge Management New York, NY, USA: Association for Computing Machinery; 2020. p. 225–234. [doi: 10.1145/3340531.3411915]
58. Abualsaud M, CHEN IX, GHAJAR K, Tahami AV, Zhang D. UWaterlooMDS at the TREC 2021 Health Misinformation Track. 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.
59. Schlicht IB, de Paula AFM, Rosso P. UPV at TREC Health Misinformation Track 2021. 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.
60. Pichel MF-, Prada-Corral M, Losada DE, C. Pichel J, Gamallo P. CiTIUS at the TREC 2021 Health Misinformation Track. 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.
61. Bondarenko A, Fröbe M, Gohsen M, Günther S, Kiesel J, Schwerter J, Syed S, Völske M, Potthast M, Stein B, others. Webis at TREC 2021: Deep Learning,

Health Misinformation, and Podcasts Tracks. The Thirtieth REtrieval Conference Proceedings (TREC 2021) National Institute of Standards and Technology (NIST), Special Publication p. 335–500.

62. Du J, Preston S, Sun H, Shegog R, Cunningham R, Boom J, Savas L, Amith M, Tao C, others. Using Machine Learning–Based Approaches for the Detection and Classification of Human Papillomavirus Vaccine Misinformation: Infodemiology Study of Reddit Discussions. J Med Internet Res JMIR Publications Inc., Toronto, Canada; 2021;23(8):e26478.

63. Cui L, Seo H, Tabar M, Ma F, Wang S, Lee D. DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining New York, NY, USA: Association for Computing Machinery; 2020. p. 492–502. [doi: 10.1145/3394486.3403092]